Grey Box Modelling of Hydrological Systems

With Focus on Uncertainties

Thordarson, Fannar Ørn; Madsen, Henrik

Publication date:
2011

Document Version
Publisher's PDF, also known as Version of record

Grey Box Modelling of Hydrological Systems

– With Focus on Uncertainties –

Fannar Örn Thordarson

Kongens Lyngby 2011
IMM-PHD-2011-XX

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

IMM-PHD: ISSN 0909-3192

Preface

This thesis was prepared at the department of Informatics and Mathematical Modelling (Informatics), the Technical University of Denmark (DTU) in partial fulfillment of the requirements for acquiring the PhD degree in engineering. The project was carried out in collaboration with DHI, through the Wellfield Optimisation project, partly funded by the Danish Strategic Research Council, Sustainable Energy and Environment Programme.

The project deals with stochastic modelling of hydrological systems with the objective of assessing the uncertainty embedded in the system formulation. The main focus is on grey box models, where the system description is formulated as a set of stochastic differential equations, but also impulse response function models are being considered.

The thesis consists of a summary report and a collection of six papers, written during the period 2007–2011. Of these, two papers have been published in conference proceedings and four research papers have been submitted for publication in international scientific journals.

Kgs. Lyngby, September 2011

Fannar Örn Thordarson

The thesis was defended on February 2nd, 2012, at the Technical University of Denmark.

The assessment committee consisted of:

- Associate Professor, Dr., Niels Kjølstad Poulsen - Department of Informatics and Mathematical Modelling, Technical University of Denmark (Chairman)
- Associate Professor, Dr., Patrick Willems - Hydraulic Laboratory, Katholieke Universiteit Leuven, Belgium (Examiner)
- Associate Professor, Dr., Michael Robdrup Rasmussen - Department of Civil Engineering, Aalborg University (Examiner)
- Professor, Dr., Henrik Madsen - Department of Informatics and Mathematical Modelling, Technical University of Denmark (Supervisor)
- Dr., Henrik Madsen - Head of Innovation, Water Resources Department, DHI, Denmark (Supervisor)

On the behalf of the Mathematical Statistics Section at DTU Informatics the defence was chaired by Associate Professor Lasse Engbo Christiansen.

Acknowledgements

First of all I would like to thank my two supervisors: At DTU Informatics, Professor Henrik Madsen for his support and guidance during the past several years, and for his positive motivation when I was confronted with various obstacles in my PhD study. At DHI, Dr. Henrik Madsen, Head of Innovation, Water Resources Department, for his helpful feedbacks and his guidance. Furthermore, I want to thank Associate Professor Peter Steen Mikkelsen from DTU Environment for his help and very useful suggestions regarding urban runoff systems.

I also appreciate the support from my colleagues at the mathematical statistics section at DTU Informatics – from fellow PhD students and postdocs to professors and secretaries. I am grateful for our interesting conversations and discussions that brought a variety of enlightenment into the everyday life at the office. I am very grateful for the cooperation with postdoc Gianluca Dorini, especially for sharing with me his expertise on groundwater management and optimisation, and his insight into modelling well fields. I am specially grateful to my fellow PhD student Anders Breinholt. Our teamwork led to several papers. Two of these are included in this thesis. Another fellow PhD student, Jan Kloppenborg Møller (now, an assistant professor at DTU Informatics) is acknowledged for sharing with me his knowledge about stochastic differential equations and stochastic modelling. In addition, I want to thank him for his very useful comments and suggestions regarding my thesis. Moreover, I want to thank all my co-authors for the collaboration that resulted in papers written and submitted during the period of the project. I am very grateful for the financial support that gave me the opportunity to acquire the PhD degree. The project was funded by the Danish Council for Strategic Research at the Danish Agency for Science, Technology and Innovation - Sustainable Energy and Environment Programme (Forsknings- og Innovationsstyrelsen; DSF - Bæredygtig

energi og miljø).

Last but not least, I am most grateful to my family and friends. Especially, I am grateful to my wife, Berglind Ósk Einarsdóttir, for her support, patience and understanding during the preparation of this thesis. My three sons, Kristófer Shawn, Dagur Örn and Askur Breki, deserve a great share of gratitude for their patience and for not losing faith in their father, despite his long periods of absensce during the last phase of the PhD study. Finally, I want to particularly thank my parents and my mother-in-law. They have been a great support to me and my family during the preparation of my thesis.

Summary

The main topic of the thesis is grey box modelling of hydrologic systems, as well as formulation and assessment of their embedded uncertainties. Grey box model is a combination of a white box model, a physically-based model that is traditionally formulated using deterministic ordinary differential equations, and a black box model, which relates to models that are obtained statistically from input-output relations. Grey box model consists of a system description, defined by a finite set of stochastic differential equations, and an observation equation. Together, system and observation equations represent a stochastic state space model. In the grey box model the total noise is divided into a measurement noise and a process noise. The process noise is due to model approximations, undiscovered input and uncertainties in the input series. Estimates of the process noise can be used to highlight the lack of fit in state space formulation, and further support decisions for a model expansion. By using stochastic differential equations to formulate the dynamics of the hydrological system, either the complexity of the model can be increased by including the necessary hydrological processes in the model, or formulation of process noise can be considered so that it meets the physical limits of the hydrological system and give an adequate description of the embedded uncertainty in model structure.

The thesis consists of two parts: a summary report and a part which contains six scientific papers. The summary report is divided into three distinct parts that introduce the main concepts and methods used in the following papers. The first part contains the basic concepts in hydrology and related hydrological models. The second part explains the grey box model by presenting stochastic differential equations and show how the equations can be linked to the available measurements. Moreover, impulse response function models are introduced as an alternative to stochastic differential equation based models, but by exploiting known hydrological models as the impulse response function in this

model makes this model framework partly physically-based. For estimating the parameters in the grey box models maximum likelihood method is used. The third important part of the summary report is predictions, and with focus on uncertainty of prediction intervals the corresponding performance measures have to include the intervals. The thesis illustrates three performance measures for this performance evaluations: reliability, sharpness and resolution. For decision making, a performance criterion is preferred that quantifies all of these measures in a single number, and for that the quantile skill score criterion is discussed in this thesis.

The second part of the thesis, which contains the papers, is divided into two different subjects. First are four papers, which consider the grey box model approach to a well field with several operating pumps. The model foundation is the governing equation for groundwater flow, which can be simplified and represented a state space form that resembles the methods used in numerical methods for well field modelling. The objective in the first two papers is to demonstrate how a simple grey box model is formulated and, subsequently, extended in terms of parameter estimation using statistical methods. The simple models in these papers consider only part of the well field, but data analysis reveals that the wells in the well field are highly correlated. In the third paper, all wells pumping from the same aquifer are included in the state space formulation of the model, but instead, but instead of extending the physical description of the system, the uncertainty is formulated to handle the spatio-temporal variation in the output. The uncertainty in the model are then evaluated by using the quantile skill score criterion. In the fourth paper, the well field is formulated by considering the impulse response function models to describe water level variation in the wells, as a function of available pumping rates in the well field. The paper illustrates, through a case study, how the model can be used to define and solve the well field management problem.

The second half of part II consists of two papers where the stochastic differential equation based model is used for sewer runoff from a drainage system. A simple model is used to describe a complex rainfall-runoff process in a catchment, but the stochastic part of the system is formulated to include the increasing uncertainty when rainwater flows through the system, as well as describe the lower limit of the uncertainty when the flow approaches zero. The first paper demonstrates in detail the grey box model and all related transformations required to obtain a feasible model for the sewer runoff. In the last paper this model is used to predict the runoff, and the performances of the prediction intervals are evaluated by the quantile skill score criterion.

Resumé

Hoved-emnet i denne afhandling er "grey box" modellering af hydrologiske systemer, samt formulering og vurdering af deres indbyggede usikkerheder. "Grey box" modellen er en kombination af en "white box" model, en fysisk baseret model, der traditionelt er formuleret som deterministiske ordinære differentialligninger, og en "black box" model, dvs. modeller baseret på statistiske input-output relationer. "Grey box" modellen består af en system beskrivelse, defineret ved et endeligt sæt af stokastiske differentialligninger, og en observationsligning. System- og observtionsligninger repræsenter tilsammen en stokastisk tilstandsrum model. I "grey box" modellen er den samlede støj opdelt i målestøj og processtøj. Processstøjen skyldes model approksimationer, uopdagede input og usikkerheder i input-serien. Estimater af process støjen kan bruges til at fremhæve manglende fit i tilstandsrumformuleringen, og yderligere støtte beslutninger for en model udvidelse. Ved brug af stokastiske differentialligninger til at formulere dynamikken i det hydrologiske system, forøger man enten kompleksiteten af modellen ved at inkludere de nødvendige hydrologiske processer i modellen, eller formulering for proces støjen kan betragtes, således at den opfylder de fysiske grænser for det hydrologiske system og giver en tilstrækkelig beskrivelse af den integrerede usikkerhed i model strukturen.

Afhandlingen består af to dele: en sammenfatning og en del som indeholder seks videnskabelige artikler. Sammenfatningen er ydeligere opdelt i tre adskilte dele, der introducerer de vigtigste begreber og metoder, der anvendes i følgende artikler. Den første del indeholder de grundlæggende begreber i hydrologi og relaterede hydrologiske modeller. Den anden del forklarer "grey box" modellen, ved at præsentere stokastiske differentialligninger og vise hvordan ligninger kan være knyttet til de tilgængelige målinger. Desuden bliver impuls respons funktion modeller introduceret, som alternativ til stokastiske

differentiallignings baserede modeller, men ved at udnytte kendte hydrologiske modeller som impuls respons funktion i denne model, bliver den delvist fysisk-baseret. Til at estimere parametrene i "grey box" modellerne anvendes maximum likelihood metoden. Den tredje vigtige del i sammenfatningen er forudsigelser, og med fokus på usikkerheder for forudsigelsesintervaller skal de tilsvarende performance mål også omfatte intervallerne. Afhandlingen illustrerer tre performance mål for vurderinger af usikkerhedsintervaller: pålidelighed, skarphed og opløsning. For beslutningstagere er et performance kriterium,der sammenfatter alle disse mål i et enkelt tal at foretrække, et sådant mål er fraktil skill scoren, der er diskuteret i denne afhandling.

Anden del af afhandlingen som indeholder artiklerne er opdelt i to forskellige fag. Først er fire artikler, der betragter "grey box" model metoden til en kildeplads med flere driftmæssige pumper. Modelfundamentet er de styrende ligninger for grundvandsstrømning, der kan forenkles og repræsenteres på en tilstandsrums form, der ligner de fremgangsmåder, der benyttes i numeriske metoder til kildeplads modellering. Målet i de første to artikler er at demonstrere, hvordan en simpel "grey box" model er formuleret og, efterfølgende, udvidet med hensyn til parameterestimering ved hjælp af statistiske metoder. De simple modeller i disse artikler betragter kun en del af kildepladsen, men dataanalysen afslører, at boringer på kildepladsen er parvis højt korreleret. I den tredje artikel, er alle boringer der pumper fra det samme grundvandsmagasin inkluderet i tilstandsrum formuleringen af modellen, men i stedet for at udvide den fysiske beskrivelse af systemet er usikkerheden formuleret til at håndtere den spatio-temporale variation i outputet. Usikkerheden i modellen er derefter vurderet ved hjælp af fraktil skill score kriteriet. I den fjerde artikel er kildepladsen formuleret ved at betragte impuls respons funktion modeller til at beskrive vandstands variationen i boringerne, som en funktion af alle tilgængelige pumpe rater i kildepladsen. Artiklen illustrerer ved hjælp af et casestudie, hvordan modellen kan anvendes til at definere og løse kildepladsstyring problemet.

Anden halvdel af del 2 består af to artikler, hvor den stokastiske differentiallignings baserede model er brugt for kloak afstrømning fra et afløbssystem. En simpel model er anvendt til at beskrive en kompleks nedbør-afstrømnings proces for et opland, men den stokastiske del af systemet er formuleret til at omfatte den stigende usikkerhed, når regnvand løber gennem systemet, samt at beskrive den nedre grænse for usikkerheden, når flow-hastigheden nærmer sig nul. Den første artikel viser i detaljer "grey box" modellen og alle relaterede forandringer der er nødvendige for at opnå en realistisk model for kloak afstrømning. I den sidste artikel anvendes denne model til at forudsige afstrømning, desuden er performance af intervallerne evalueret ved hjælp af fraktil skill score kriteriet.

List of publications

Papers included in the thesis

[A] Fannar Örn Thordarson, Henrik Madsen & Henrik Madsen (2009) Grey box modelling of a groundwater well field. In proceedings: *ModelCare 2009, 7th International Conference on Calibration and Reliability in Groundwater Modeling, Managing Groundwater and the Environment*, pp. 67-70. September 20-23, Wuhan, China.

[B] Fannar Örn Thordarson, Henrik Madsen & Henrik Madsen (2010) Predictions for groundwater well fields using stochastic modelling. In proceedings: *HydroPredict 2010, 2nd International Interdisciplinary Conference on Predictions for Hydrology, Ecology and Water Resources Management*. September 20-23, Prague, Czech Republic.

[C] Fannar Örn Thordarson, Henrik Madsen, Gianluca Dorini & Henrik Madsen (2012) Stochastic well field modelling using the grey box approach. Submitted to *Advances in Water Resources*.

[D] Gianluca Dorini, Fannar Örn Thordarson & Henrik Madsen (2011) Stochastic simulation and robust optimal management of well fields using impulse response function models. Submitted to *Water Resources Research*.

[E] Anders Breinholt, Fannar Örn Thordarson, Jan Kloppenborg Møller, Peter Steen Mikkelsen, Morten Grum & Henrik Madsen (2011) Grey box modelling of flow in sewer systems with state dependent diffusion. *Environmetrics*, **22**(8):946-961.

[F] Fannar Örn Thordarson, Anders Breinholt, Jan Kloppenborg Møller, Peter Steen Mikkelsen, Morten Grum & Henrik Madsen (2012) Evaluation of probabilistic flow predictions in sewer systems using grey box models and a skill score criterion. Accepted for publication in *Stochastic Environmental Research and Risk Assessment*.

Other Publications

The following papers were also prepared during the project period. The scientific content is covered, or partly covered by the included papers. Therefore, these papers are not included in the monograph.

- Gianluca Dorini, Fannar Örn Thordarson, Peter Bauer-Gottwein, Henrik Madsen, Dan Rosbjerg & Henrik Madsen (2011) A convex programming framework for optimal and bounded sub-optimal well field management, considering pump settings and water distribution networks. Submitted to *Water Resources Research*.

- Gianluca Dorini, Fannar Örn Thordarson, Henrik Madsen & Henrik Madsen (2011) Analysis and treatment of the Søndersø time series - grey box well field modelling. *IMM-Technical Report-2011-04*. DTU Informatics. Kgs. Lyngby, Denmark.

- Anders Breinholt, Morten Grum, Henrik Madsen, Fannar Örn Thordarson & Peter Steen Mikkelsen (2012) Informal uncertainty analysis (GLUE) of continuous flow simulation in a hybrid sewer system with infiltration inflow - consistency of containment ratios in calibration and validation?. Submitted to *Hydrology and Earth System Sciences*.

- Anders Breinholt, Fannar Örn Thordarson, Jan Kloppenborg Møller, Morten Grum, Peter Steen Mikkelsen & Henrik Madsen (2012), Identifying the appropriate physical complexity of stochastic gray-box models used for urban drainage flow prediction by evaluating their point and probabilistic forecast skill. Submitted to *Water Resources Research*.

Additionally, the following paper was also prepared during the project period. However, the scientific content of the paper is not relevant to this thesis and, thus, it is not included:

- Fannar Örn Thordarson, Henrik Madsen, Henrik Aalborg Nielsen & Pierre Pinson (2010) Conditional weighted combination of wind power forecasts. *Wind Energy*, **13**(8):751-763.

Contents

Part I

Summary Report

CHAPTER 1

Introduction

The purpose of hydrological models is to understand the behaviour of hydrological systems, where the focus is on generating predictions for controlling and managing water resources so that human lives are protected and property damages are prevented. Over the last decades the hydrological society has been confronted with new challenges that involve an increased focus on extreme events and water quality. This calls for new modelling approaches in order to gain more information about future scenarios that describe uncertainties for both short-term and long-term predictions. The model approaches have either been deterministic, where the physical laws are the foundation in the mathematical framework, or purely statistical ones where available data is employed to build a model exclusively based on the input-output relation.

1.1 Motivation

The main objective of the thesis is to combine these two different approaches. This is carried out by formulating the model such that it includes the significant dynamic behaviour of the physically-based model, but enables statistical methods and tools for estimating the model parameters and assessing the uncertainties in the model structure. Hydrological systems are most often highly complex where usually - due to the law of conservation - many interconnected

processes are needed in order to describe the system evolution in both time and space. Hydrological processes are physical phenomena, where the system dynamics are often best formulated in continuous time by considering ordinary differential equations, or partial differential equations. In order for hydrological models to efficiently describe the dynamics in a deterministic hydrological system, a thorough understanding of the system and all influential subprocesses is required. This is obtained by a rather detailed description of all the processes involved. Consequently, the number of parameters that need to be estimated in the model is typically large. A mathematical framework that is based on such a formulation is very deterministic, and often referred to as white box models.

However, in real world applications the ideal hydrological model, where the system and all subprocesses are well described, does not exist. All hydrological models are only approximations of the true process, but the model is considered to adequately describe the system behaviour when the residual series, the difference between the model predictions and the measurements, is minimised and observed as a series of white noise terms. If the deterministic model does not include all the necessary influential factors, the residual series will render a systematic pattern, and cause the model to depart from the measured output. The model will then become incompetent for its purpose. Detailed white box hydrological models are usually applied for simulation purposes, where the objective is to determine the long-term effects of the system response on predefined input sequences.

On the other hand, statistical models are desirable for short-term predictions, since the statistical methods make it possible to use rigorous stochastic dynamical models that provide a measure of the inherent uncertainty for the model predictions. However, statistical models are discrete time models that do not normally contain any physical knowledge regarding the system, and the physical parameters are partly hidden in the discrete time parameterisation. Thus, for long-term predictions of physical phenomena, which is solely based on statistical models, the model output is not adjusted towards the physical drift embedded in the physical knowledge of the system. Due to the lack of physics in the model structure, where only the input-output data and statistical methods are used to formulate the model, statistical models are called black box models.

To maintain the physical interpretation of the model, it would be suitable to use formulation and apply an estimation method, where the parameterisation is kept in continuous time. The model proposed in the thesis is based on the most important physical knowledge of the hydrological process, but includes an additional stochastic term to cope with uncertainties in the model formulation and in the observations. The parameters in the model are physically interpretable and estimated by applying statistical methods. The model ap-

proach is called grey box model, since the basic model structure is inherited from the white box models, usually in the form of ordinary differential equations, but the parameter estimation and the uncertainty assessment is obtained using statistical methods. The grey box concept is illustrated in the diagram in Figure 1.1, showing the contributions from the white box and the black box approaches.

Considering the proposed grey box approach, the recommendation is to adapt a simple model to describe the system dynamics. The reason for this recommendation is that the law of parsimony tells us that the simplest adequate models are preferred in order to obtain a model and parameters that are identifiable from data. Such simple model can then be extended, based on the estimation results, where the parameter estimates and their variances play a central role in identifying the lack of fit in the system formulation. The parameter estimation also contains estimation for the uncertainty of the parameters. This indicates that an improved model structure can be obtained by considering the lack of fit in the deterministic part of the model, where extensions call for more deterministic equations in order to remove the unwanted uncertainty. On the other hand, the uncertainty can also be formulated in accordance with the knowledge regarding the hydrological system. The former method has been applied in many fields, e.g., chemical engineering (*Kristensen et al.*, 2004a), dynamic models for air temperature (*Søgaard*, 1993) and heat dynamics of buildings (*Bacher and Madsen*, 2011). However, the latter approach has not received much attention until recently (*Møller et al.*, 2010a,b, *Philipsen et al.*, 2010). This thesis is dedicated to the uncertainty part of the grey box model, where simple models are adopted to represent the dynamic behaviour of the hydrological system. The uncertainty is formulated to obtain reasonable prediction intervals for the model output. Thus, the grey box approach provides adequate and operational models for the system. In that way the models are not only physically interpretable, but do also depend on real time measurements and, therefore, useful for both short-term and long-term predictions in connection with online control and optimisation.

The grey box model is considered in connection with two different areas within hydrology. Firstly, in connection with well field modelling, i.e., modelling the pressure heads in several operating wells in a well field, and, secondly, in connection with sewer runoff modelling for a drainage system where a combined flow of rainwater and wastewater is diverted from a catchment.

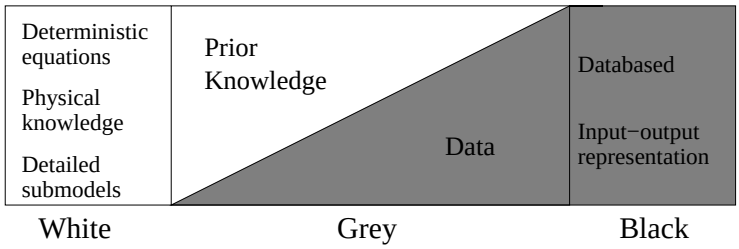

Figure 1.1: The grey box modelling concept. The prior knowledge from the physical structure in the white box approach and the available data from the measured input and output variables applied to the black box model, is combined in a grey box modelling approach.

1.2 Overview of the thesis

The thesis is divided into two parts; a summary report and a collection of papers that have been written or prepared during the period of my PhD study. The summary report is written as an introduction to the methods and models that have been applied in the included papers. The thesis is based on six papers; two conference papers and four research papers, submitted or accepted for publication in international journals.

The summary report includes five chapters. Following this introduction, three main topics are explained in the three subsequent chapters. In Chapter 2 some basic and useful concepts in hydrology are presented, along with the essential physical structure for both the groundwater flow model and sewer runoff model.

Chapter 3 covers the mathematical theory behind the models applied in the papers, i.e., the theory of stochastic differential equations and models based on impulse response functions. This is followed by an introduction to the maximum likelihood method. The method is applied for all six papers in the parameter estimation. The stochastic differential equations consist of a drift term, corresponding to an ordinary differential equation for maintaining the physical knowledge in the equation, and a diffusion term accounting for the stochasticity in the equation. The focus area of my thesis is the uncertainties in the system structure. Hence, modelling with stochastic differential equations, and in particular the diffusion term is given special attention in the chapter.

An overview of the topics used for prediction and uncertainty assessment are contained in Chapter 4. To evaluate the uncertainty, the measurement criteria for reliability, sharpness and resolution are described. This is followed by the skill score criterion. It is considered as a global measurement for evaluating

prediction intervals, and more appropriate as an overall performance measure to distinguish between model proposals.

Chapter 5 contains discussion and some concluding remarks regarding both the summary report and the included papers. Subsequently, some future perspectives are presented regarding the models proposed in the papers.

Following the summary report are the six papers, where Papers A to D are dedicated to well field modelling, and Papers E and F deal with modelling and prediction for the drainage sewer runoff.

CHAPTER 2

Hydrology

Life depends on water. Our entire way of life is based on accessibility to water resources and water abundance. In order to be able to harvest and produce food, people have for thousands of years established societies and cultures around water resources.

Hydrology is the study of water, with focus on movement, distribution and quality of water. Hydrology is divided into several domains, where each domain has its own environmental identity, including hydrometeorology, surface hydrology, hydrogeology, management of drainage basins and assessment of water quality.

2.1 The hydrological cycle

To understand the movement of water in our environment, a fundamental step is to acknowledge how water circulates in the world in which we live. To establish the origin of the rainfall and the source of the lakes and rivers that never stop flowing down the hills, we have to realise that water circulates both in the atmosphere and below Earth's surface.

The circulation is named the hydrological cycle, and is sketched in Figure 2.1. In its simplest form, this is described as the movement of water, as it evapo-

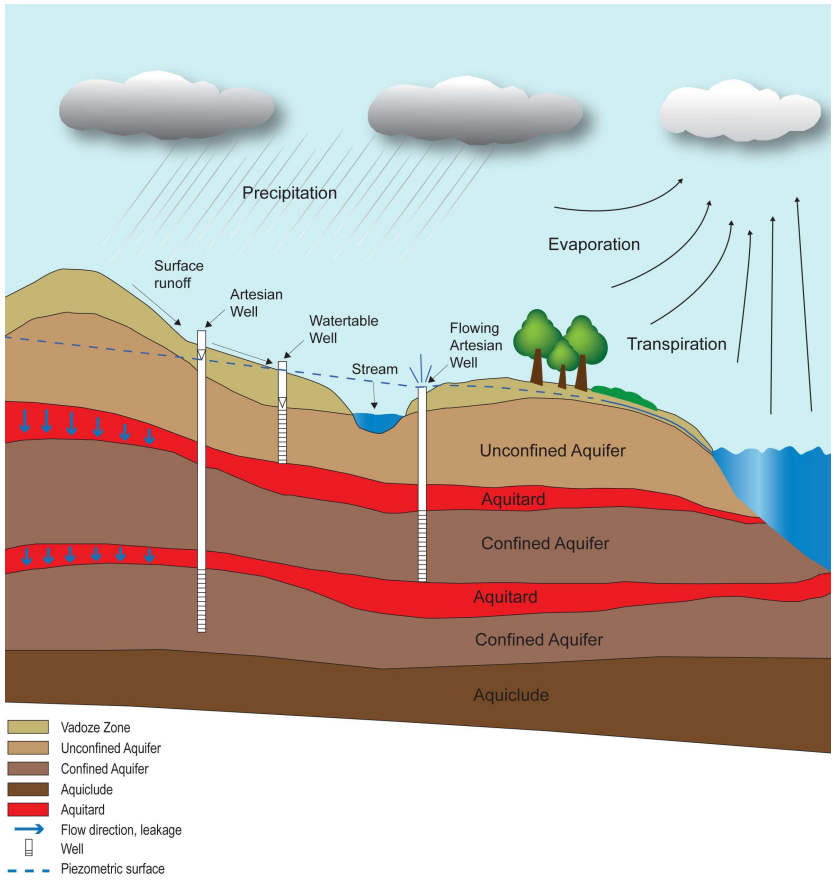

Figure 2.1: The hydrological cycle, including the concepts in hydrogeology.

rates from the surface of both sea and land to the atmosphere. Water vapour is transported in the atmosphere until it condenses (clouds) and, subsequently, may dissolve over land in the form of precipitation. The precipitation that falls on the surface area is partly collected into streams and rivers by surface runoff that eventually runs back into the sea. The remaining part of the rainfall, however, is infiltrated into the soil.

The infiltrated water is stored in the subsurface, also called vadose zone. The area below land surface is roughly divided into the subsurface and the groundwater. You distinguish between the two parts by means of the watertable, which is an elevation of saturation for the water stored below surface. From the soil in the vadose zone the water flows as a subsurface flow into streams and lakes, or by gravitational forces percolates further into the ground and

recharges the groundwater. The groundwater flow diverts the water towards lakes, streams and the ocean. The precipitated water can also be intercepted by vegetation and, subsequently, return to the atmosphere by the so-called transpiration; a process similar to evaporation, but assigned to loss of water vapour from plants, flowers, etc.

However, the cycle is not that simple. Each state in the cycle is faced with stages that cannot be overlooked, e.g., the precipitation can fall anywhere, also directly into the sea during or right after cloud formation. The time of the cycle is not uniform. In other words, during droughts it seems it will never rain, whereas during floods it seems the rain will never stop. Also, the intensity of the precipitation events depends on climate and geographical location.

2.2 Hydrological modelling

Hydrological models are a simplified descriptions of parts of the hydrological cycle. The purpose is to gain such information about the hydrologic process that it can be used for predictions for states of the hydrological system. The objective in hydrological modelling is essentially to determine a description for the flow as it passes from the input to the output, i.e., to obtain an acceptable input-output representation for the hydrological system. The methods of flow routing depend on knowledge about storage capacities in the hydrological system and, in general, either deterministic models or stochastic models are used to evaluate the storage.

As explained in Chapter 1 the deterministic models (or the white box models) are very detailed descriptions of the hydrological system, and are usually based on physical knowledge of the system dynamics, only. The system parameters are obtained from hydrological surveys related to the system's characteristics. Thus, they are predefined in the model structure without any uncertainty. The deterministic models tend to become fairly complicated, because the model accuracy is optimised by improving the system with inclusion of as many processes and subprocesses as possible in order to minimise the output error. In contrast, stochastic hydrological models refer to statistical models (or black box models), i.e., the model structure is obtained by correlating the available input and output data series for the hydrological system in question. Processes in hydrology that are categorised as black box models are usually models with the primary goal of making short-term predictions, e.g., rainfall-runoff and flood forecasting.

To accomplish better predictions and, consequently, improved water resources management systems, the mainstream in modern hydrological modelling is

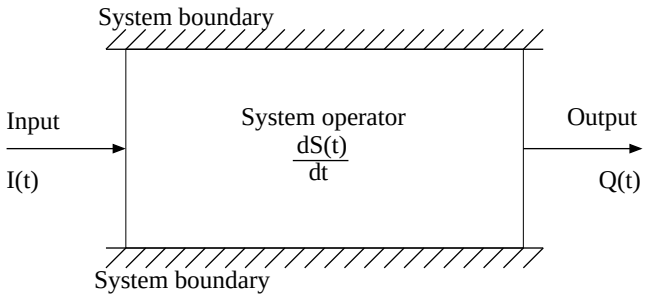

Figure 2.2: Representation of a closed system. Changes in the system operator, the storage S, at time t are due to the difference between the input I and the output Q.

focused on gaining a more general understanding of the behaviour of the hydrological systems.

2.2.1 Storage equation

The fundamental requirement in hydrological modelling is that the water balance in the hydrologic process is preserved in such a way that within a closed system the same quantity of water exiting the water system is the same as the quantity entering the system. Changes in the system are then illustrated as the difference between the input and the output at some specific time. In continuous-time these system changes are described with the continuity equation (*Douglas et al.*, 2001) [1] which states, as previously described, that a flux of water going into the system at the input must emerge at the output. The law of mass conservation is illustrated in Figure 2.2.

In a storage $S(t)$ with the input $I(t)$ and corresponding output $Q(t)$, and which is not influenced by alternative external factors (e.g., subprocesses in the form of lateral flow to a river), the storage equation, for time t, is written

$$\frac{dS(t)}{dt} = I(t) - Q(t). \tag{2.1}$$

If a difference is detected between the input and the output at the time t, the system is considered to be unsteady, i.e., $I(t) \neq Q(t) \Leftrightarrow dS(t)/dt \neq 0$. Furthermore, the flow in the storage is considered to have a constant density. If the density is varying the flow is compressible, but compressible flows are not being considered in the hydrological studies in this thesis.

[1] Also called the law of mass conservation

The differential equation (2.1) is a traditional deterministic representation of flow through a storage, simply stating that all influential processes are well described within the system operator (Figure 2.2). This indicates that all sub-processes and potential input variables are included in the system formulation. However, this is very seldom the case since hydrological processes are complex phenomena. Due to the system complexities, it is usually very difficult to conclude that a system has reached a complete description. The available data for the system usually contains measurements for the input and the output variables, but unavoidably a deviation is exposed as the model is compared to the measured output. Even for the "perfect" model, a minor discrepancy is detected, but this discrepancy increases with the lack of a description for the system dynamics. Therefore, a noise term is added to the system representation in (2.1) to account for the deviation between the input and the output of the system, i.e.,

$$\frac{dS(t)}{dt} = I(t) - Q(t) + \text{"Noise"}. \tag{2.2}$$

2.3 Surface water hydrology

Surface water hydrology refers to the theory of movement of water on land surface. Flowing water on Earth's surface is a vital part of the hydrological cycle. A surface runoff is a derivation of precipitation fallen on land areas, where the excess water is collected into lakes and rivers that diverts the water from the rainfall back to the sea. Incidents linked to surface water hydrology are directly observable – both those that are due to natural causes and those that are man-made structures created to avoid nuisance. In urban areas, derivation of all excess water is of great importance, since overflows in these areas can severely affect properties and human lives as well as causing pollution of water supplies and disruption of communication and transport.

For a particular rainfall-runoff system, catchment is the land area that receives the water from the rainfall, whereas the land area that contributes to the surface runoff to the catchment outlet is called watershed. The relation between rainfall and corresponding runoff has been studied for decades. A fundamental tool to visualise this particular rainfall-runoff correlation is a hydrograph. A hydrograph shows how the flow rate evolves in time for a given location in the rainfall-runoff system.

2.3.1 The unit hydrograph

As a response to a rainfall, a unit hydrograph can be used to illustrate an outlet flow from a watershed or a catchment. It is usually illustrated graphically, and is basically an impulse response function of a linear time-invariant system (*Madsen*, 2008), showing how a discharge from a watershed evolves in time, subsequent to a single unit of rainfall on the catchment (*Sherman*, 1932). The unit hydrograph can be visualised as illustrated in Figure 2.3, by setting r and s equal to one, and t is a single time-step. The unit hydrograph is considered unique for a given watershed, where several terms can be identified to characterise the watershed. One of these terms is the time delay t_p from the time of the rainfall to the time of the corresponding hydrograph to reach its maximum flow (peak flow), shown in Figure 2.3. This time delay is often referred to as retention time and plays a central role in characterising rainfall-runoff flows. However, several requirements have to be fulfilled for the unit hydrograph analysis (*Chow et al.*, 1988): The duration of the rain event has to be brief, and the catchment should be small. Also, the rainfall is assumed to be uniformly distributed through the entire drainage area. For a single catchment, an increase in the rainfall causes an increase in the hydrograph as well, such as displayed in Figure 2.3.

For long periods of dry-weather situations, the only contribution to the measured runoff is the so-called baseflow. Several methods have been proposed to separate the baseflow from the direct runoff flow, e.g., the recession curve approach and the arbitrary approach (*Gupta*, 2008, *Jonsdottir et al.*, 2006a). For a flow in a sewer drainage system, the baseflow corresponds to the wastewater from the household in the catchment, and is usually seasonally observed due to the traditional daily and weekly behaviour of the household.

2.3.2 Runoff models

One of many lumped flow routing methods that have been developed (*Chow et al.*, 1988, *Viessman and Lewis*, 1996) considers the hydrological system as a series of linear reservoirs. Reservoir is referred to as linear when the reservoir storage is in linear relation to the output flow from the reservoir, linked with a storage coefficient k [T] that represents the retention time for the flow through the reservoir, i.e.,

$$S(t) = kQ(t) \Leftrightarrow Q(t) = \frac{1}{k}S(t). \tag{2.3}$$

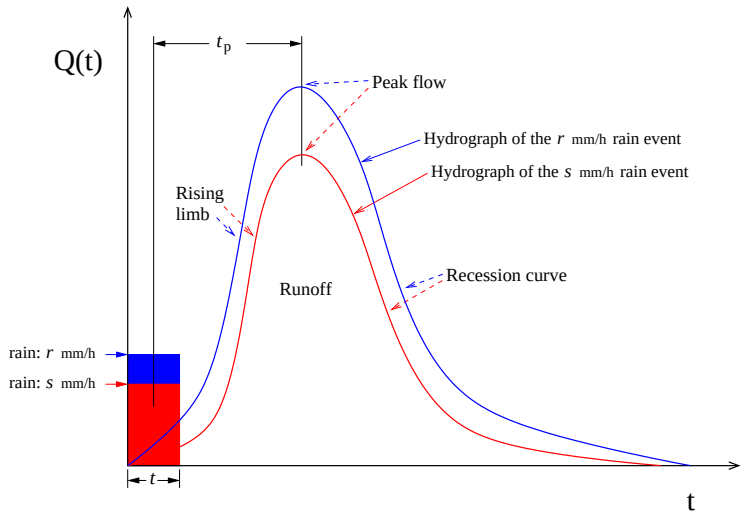

Figure 2.3: Two hydrographs for a single uniform rain event falling on a catchment. A rain of s mm/h results in a runoff flow as shown by the red hydrograph. An increase in rain from s to r mm/h will influence the size of the peak flow as the hydrograph is increased as well (from the red curve to the blue curve, respectively).

Thus, the differences in the output flow from the reservoir is due to the difference in storage in the reservoir:

$$dQ(t) = \frac{1}{k} dS(t),$$

and for changes due to time, the time derivative for the output is written

$$\frac{dQ(t)}{dt} = \frac{1}{k} \frac{dS(t)}{dt}. \tag{2.4}$$

By replacing the time derivative of the storage, dS/dt, in (2.4) with the storage equation (2.1), the time derivative for the output flow in a single reservoir becomes

$$\begin{aligned} \frac{dQ(t)}{dt} &= \frac{1}{k} \left(I(t) - Q(t) \right) \\ &= \frac{1}{k} I(t) - \frac{1}{k} Q(t). \end{aligned} \tag{2.5}$$

Using the condition $Q(0) = 0$ and considering the input as an unit impulse, the solution for the single reservoir output flow is

$$Q(t) = \frac{1}{k} e^{-\frac{t}{k}}. \tag{2.6}$$

Hence, the solution is an impulse response function with an exponential decay. The response time to the unit impulse, T for a single reservoir has exponential distribution with rate $1/k$, with mean $\mathrm{E}\{T\} = k$ and variance $\mathrm{V}\{T\} = k^2$. This corresponds to the top hydrograph on the right side of Figure 2.4.

By considering a system of N reservoirs (displayed on the left side of Figure 2.4) the results from the single reservoir flow equation (2.6) are utilised, and also the fact that the retention time in reservoir n, for $n = 1, \ldots, N$, is assumed independent. The resulting impulse response function for the flow through N linear reservoirs in a series, where the storage parameter k is the same for all reservoirs (*Nash*, 1957), is

$$Q_N(t) = \frac{1}{k\Gamma(N)} \left(\frac{t}{k}\right)^{N-1} e^{-\frac{t}{k}}, \tag{2.7}$$

which is a gamma distribution where N represents the shape parameter and the rate parameter is the inverse of the retention time k, i.e., $1/k$. The mean and variance for flow through the N reservoirs is, due to the gamma distribution, $\mathrm{E}\{T_N\} = kN$ and $\mathrm{V}\{T_N\} = k^2 N$, respectively.

The system of linear reservoirs can be presented directly from Eq's. (2.2) and (2.3) on a state space form, where the nth state, S_n for $n = 2, \ldots, N$, is described as

$$\frac{dS_n(t)}{dt} = \frac{1}{k} S_{n-1}(t) - \frac{1}{k} S_n(t) + \text{“Noise”}.$$

If the storage parameter is the same for all states, the mean of the gamma distribution in (2.7) can be adopted in the state space formulation to include a parameter for the mean retention time for the flow through all N states. Hence, $k = T_N/N$ and the state space formulation is written

$$\frac{d}{dt}\begin{bmatrix} S_1(t) \\ S_2(t) \\ \vdots \\ S_N(t) \end{bmatrix} = \begin{bmatrix} -\frac{N}{T_N} & 0 & \cdots & 0 \\ \frac{N}{T_N} & -\frac{N}{T_N} & & \\ \vdots & \ddots & \ddots & \\ 0 & & \frac{N}{T_N} & -\frac{N}{T_N} \end{bmatrix} \begin{bmatrix} S_1(t) \\ S_2(t) \\ \vdots \\ S_N(t) \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} I(t) + \textbf{“Noise”}. \tag{2.8}$$

The state space description can readily be extended to the case of different storage parameters for the individual states.

2.3.3 Drainage systems

Drainage is a term that applies to the process of removing excess water from catchments so as to prevent overflow and in that way protecting properties and

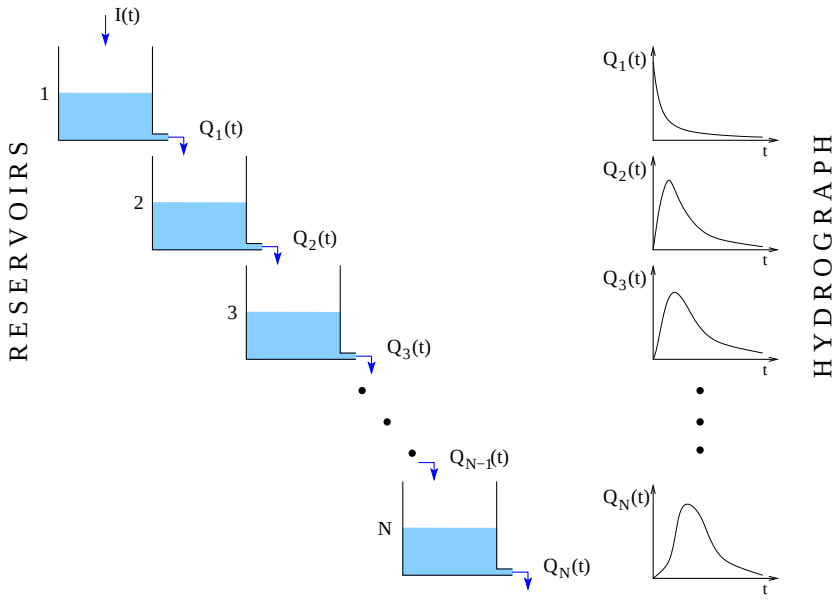

Figure 2.4: A system of N linear reservoirs. Corresponding hydrographs for the output $Q_n(t)$ are shown to the right.

lives. For rural areas, or areas that have not been developed, the drainage occurs naturally as a part of the hydrological cycle (Figure 2.1) and infiltrates into the vadose zone. For developed areas, however, or so-called urbanised areas, the human factor has severely influenced the drainage of the excess water in the catchment.

Urban drainage was introduced in order to improve sanitary conditions in the populated areas, and to divert the flow out of these areas. In order to remove both wastewater and rainwater, and thereby minimise the inconvenience for the population, pipe networks are constructed below ground surface in cities and towns. In cities the rain falls on either a permeable or impermeable area. The permeable areas drain the water to the subsurface by infiltration, but in the impermeable areas the excess water is collected from the paved areas, e.g., roofs and streets, by open channels linked to the drainage systems. Furthermore, through wastewater from the households the population also contributes to the runoff system.

In papers E and F the linear reservoir model in (2.8) is applied to a simple model of urban drainage system.

2.4 Groundwater hydrology

Groundwater hydrology deals with occurance, movement and quality of water stored in the saturated part of the underground zone. A geologic formation in the saturated zone capable of storing a significant amount of groundwater is known as an aquifer; i.e., a sediment that yields water in quantities that are sufficient for a well or a spring (see Figure 2.1). The hydraulic conductivity of an aquifer is high, meaning that the aquifer – within a reasonably short time – is capable of transferring water from one location to another within the aquifer. The zone of saturation is usually categorised into several layers of different aquifers. These are vertically separated by either layers of much lower permeability than the nearby aquifers, referred to as aquitards, or by layers that are almost impermeable, the so-called aquicludes. They form flow barrier between aquifers.

An aquifer can be considered as either confined or unconfined. An unconfined aquifer is an aquifer that is directly influenced by the vadose zone, i.e., the aquifer is not bounded by an aquitard or an aquiclude on top. For multiple layers of aquifers, the upper most aquifer is always defined as an unconfined aquifer, with the varying watertable as its upper limit and, thus, is recharged by rain or irrigation water that percolates from the Earth's surface through the vadose zone. The water level in a bore hole that is drilled into an unconfined aquifer is the same as the watertable, separating the saturated and unsaturated zones.

The confined aquifer, however, has an aquitard as an upper barrier, and an aquitard or an aquiclude below. The level of the watertable of the recharge area of the confined aquifer is usually much higher than the top of the confined aquifer itself, indicating that the water in the confined aquifer is pressurised. A bore hole that is drilled into such an aquifer has a water level that rises significantly above the top of the aquifer.

In Denmark, the main resource for drinking water is the groundwater, which is transported from pumping wells – one or several in a region often referred to as a well field – to the consumers.

2.4.1 Darcy's law

The basic equation of groundwater flow is Darcy's law, an analogue to Fourier's law for heat transfer, and Ohm's law in electrical circuits. The equation is named after Henri Darcy, who in 1856 experimented with flows in a pipe filled

with sand. He discovered that the flow rate Q in the pipe was inverse proportional to the length of the sand filter L and proportional to both the cross-sectional area of the pipe A and the head drop between the start and the finish of the filter, Δh. With an addition of a constant of proportionality for the sand, the so-called hydraulic conductivity κ, Darcy's law is

$$Q = \kappa A \frac{\Delta h}{L}, \tag{2.9}$$

and in general describes water transport in a porous material. As mentioned above, Darcy's law describes one-dimensional flow, only. The flow equation can be generalised so as to apply to three-dimensional flow, and that leads to the governing equation for groundwater flow (see, e.g., *Gupta*, 2008).

Darcy's law cannot be generalised for all flows. It is only valid for laminar flows, and for flows in Newtonian fluids. Furthermore, for flows through extremely fine-grained material, Darcy's law does not apply, nor does it if the medium is not fully saturated.

2.4.2 Stochastic groundwater flow

In a confined anisotropic aquifer the governing equation for groundwater flow is a partial differential equation:

$$S_s \frac{\partial h}{\partial t} = \nabla \cdot \boldsymbol{\kappa} \nabla h + R \tag{2.10}$$

where h [L] is the hydraulic head, $\boldsymbol{\kappa}$ [LT^{-1}] is the tensor matrix of the hydraulic conductivity, S_s ([L^{-1}]) is the specific storage and R [T^{-1}] represents any external stress affecting the groundwater flow.

To obtain a successful model for the groundwater flow, where water is discharged from the aquifer at several locations simultaniously, the groundwater flow equation (2.10) has to be discretised and solved numerically. This is done by dividing the well field into a number of cells. This is the general numerical methodology for partial differential equations, where usually the finite element or finite difference methods are applied to solve the equation. In hydrogeology several commercial softwares have been developed for simulation of groundwater, e.g., MODFLOW (*McDonald and Harbaugh*, 1983) and MIKE-SHE (*Madsen et al.*, 2008).

In the thesis, the cell is assumed to be of an arbitrary form and the cells are discretised without taking restrictions related to the shape of the cells into consideration. For that the finite volume method is applied (*Rozos and Koutsoyiannis*,

2010). Hence, Eq. (2.10) is integrated with respect to the volume V for the discretised cell in the well field. Consequently, assuming that the input stresses R and the specific storage is space invariant for each cell, the divergence theorem (*Adams*, 1999) can be applied to obtain

$$S_s V \frac{\partial h}{\partial h} = \int_{\boldsymbol{S}} \boldsymbol{\kappa} \nabla h \cdot \boldsymbol{n} d\boldsymbol{s} + RV \tag{2.11}$$

where the integral on the right hand side is a surface integral of the total discharge through the surface $\boldsymbol{S}$, surrounding the volume V. The vector $\boldsymbol{n}$ is a unit vector, normal to the surface element $d\boldsymbol{s}$ and pointing outwards from the volume.

The surface integral (2.11) cannot be solved analytically and numerical methods have to be used. An assumption to simplify the calculation of the integral is to consider equipotential lines (no-flow lines) at the edges of the discretised cell. Consequently, $h \cdot \boldsymbol{n}$ is equal to the gradient of h along the marginal of the cell. Thus, the surface integral can be reduced to an aggregation of Darcian fluxes to the cell from all the neighbouring cells (*Anderson and Woessner*, 2002).

To calculate the water level in cell i (h_i) under influence from the J neighbouring wells (see Figure 2.5), the integral in Eq. (2.11) is simplified to

$$S_{s,i} V_i \frac{dh_i}{dt} = \sum_{j=1}^{J} \frac{\kappa_{ij} \bar{A}_{ij}}{D_{ij}} \left(h_j - h_i\right) + R_i V_i \tag{2.12}$$

where D_{ij} is the distance between wells i and j, and $\bar{A}_{ij}$ is the cross-section area between the same cells, with conductivity κ_{ij}. Furthermore, V_i, $S_{s,i}$ and R_i are the same as mentioned previously, but are now related to cell i only. Assuming that the cross-section area is a constant, is only valid for confined aquifers. For unconfined aquifers, Eq. (2.12) can also be utilised, but then the cross-sectional area $\bar{A}_{ij}$ is a function of the water head difference in the cells, and the groundwater flow equation becomes nonlinear. However, unconfined aquifers are not of interest in the following studies and are, therefore, not dealt with.

For a homogeneous isotropic confined aquifer the aquifer thickness b is considered to be uniform. Consequently, several assumptions can be attained regarding the parameters in the groundwater flow:

$$A_i = \frac{V_i}{b}, \quad S_i = S_{s,i} b, \quad T_{ij} = \kappa_{ij} b \quad \text{and} \quad E_{ij} = \frac{\bar{A}_{ij}}{b}$$

where A_i [L^2] is the base area of cell i, S_i is its storage coefficient, or storativity [-], T_{ij} corresponds to the transmissivity of the flow between the cells i

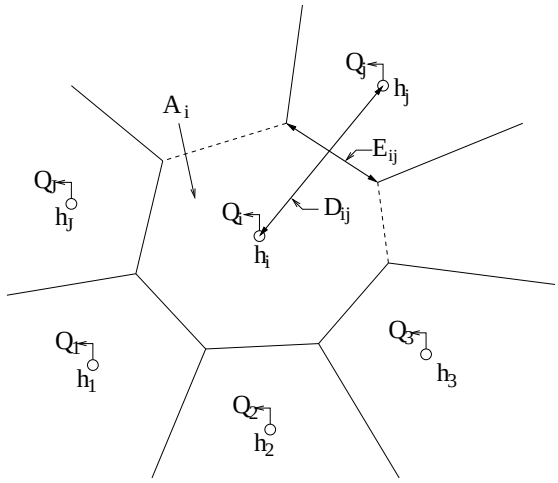

Figure 2.5: A sketch of cell i with an operating well included. Also included, the parameters related to cell i and its coupling to the neighbouring cell j.

and j [L^2T^{-1}], and E_{ij} is the screening of the cross-sectional flow [L]. The flow equation for cell i is then written as

$$S_i A_i \frac{dh_i}{dt} = \sum_{j=1}^{J} \frac{T_{ij} E_{ij}}{D_{ij}} (h_j - h_i) + W_i \tag{2.13}$$

where $W_i = R_i b_i A_i$ and accounts for sources and sinks that either discharge or recharge cell i to maintain the water balance in the system.

In well field modelling the discharge of water is considered to be the most influential sink that affects the aquifer, and for a cell with a pumping well included the pump rate of the well has to be subtracted from the groundwater flow equation for that particular cell. However, due to the storage coefficient, which represents the porousity of the aquifer, the drawdown in the cell is not as rapid as the one detected inside the well where pure water is discharged. Hence, the water level outside the well is less affected by the pumping than the water level inside. Thus, the discharge rate is multiplied by the storage coefficient in order to calculate the water drawdown in the cell.

The source/sink term in the groundwater flow equation (2.13) can then be written

$$W_i = -S_i Q_i + L_i A_i (H_0 - h_i) \tag{2.14}$$

where Q_i is the positive discharge flow through the well in the cell, L_i is the

leakage coefficient through the aquitard, and H_0 corresponds to the piezometric surface in the cell. The linear equation for the water drawdown in cell i in a confined aquifer is then written

$$\frac{dh_i}{dt} = \frac{1}{S_i A_i} \sum_{j=1}^{J} \frac{T_{ij} E_{ij}}{D_{ij}} (h_j - h_i) + \frac{L_i}{S_i} (H_0 - h_i) - \frac{1}{A_i} Q_i. \tag{2.15}$$

The stochastic groundwater model considered in Papers A, B and C is presented on a state space form. In its simplest formulation each state contains a single operating well, only. Hence, for a well field with N wells, the proposed groundwater model for the well field is described by only N states in the system description. By such a simple formulation of the complex physical system as an aquifer, the model structure is a rough approximation of the real system. However, the discrepancy from reality for each of the states can be quantified by adding a noise term to the states, and an estimation of the noise term will provide a measure of uncertainty of the model.

A stochastic groundwater model with N wells is then written on a state space form as

$$\begin{aligned} \frac{d}{dt} \begin{bmatrix} h_1 \\ \vdots \\ h_N \end{bmatrix} = & \begin{bmatrix} -\frac{1}{S_1} \left(\frac{1}{A_1} \sum_{j=1}^{J} \frac{T_{1j} E_{1j}}{D_{1j}} + L_1 \right) & \cdots & \frac{T_{1N} E_{1N}}{S_1 A_1 D_{1N}} \\ \vdots & \ddots & \vdots \\ \frac{T_{N1} E_{N1}}{S_N A_N D_{N1}} & \cdots & -\frac{1}{S_N} \left(\frac{1}{A_N} \sum_{j=1}^{J} \frac{T_{Nj} E_{Nj}}{D_{Nj}} + L_N \right) \end{bmatrix} \begin{bmatrix} h_1 \\ \vdots \\ h_N \end{bmatrix} \\ & + \begin{bmatrix} \frac{L_1}{S_1} & -\frac{1}{A_1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{L_N}{S_N} & 0 & \cdots & -\frac{1}{A_N} \end{bmatrix} \begin{bmatrix} H_0 \\ Q_1 \\ \vdots \\ Q_N \end{bmatrix} + \textbf{"Noise"} \end{aligned} \tag{2.16}$$

and corresponds to the model approach introduced in Paper C.

2.5 Concluding remarks

In this chapter an overview of some fundamental and widely used concepts in hydrology have been introduced. The hydrological cycle has been described, and parts of it are included in this monograph. Firstly, the sewer runoff. The modelling approach proposed in Papers E and F is an extension of the basic model for a series of linear reservoirs applied to urban drainage systems. The runoff/drainage model is presented on a state space form to account for the retention time in the system. Secondly, the groundwater flow equation is simplified so that by assumptions it is transformed from a partial differential equation to a set of ordinary differential equations on a state space form.

In both cases the modelling approach involves several approximations before arriving at a simplified description of the system. The aim of the simplification is to obtain a structure that describes the flow from the input to the output on a state space form where each state is presented by an ordinary differential equation. However, for each step towards the simplified model, the embedded uncertainty in the model structure is increased; an uncertainty that has to be taken into account in order to obtain a reasonable assessment of the variation of the model output. A description of the uncertainty is also needed in order to account for the fact that in many cases the forcing or input to the system is not known exactly.

In the following chapter the additive noise in the differential equation is described. This leads to a set of stochastic differential equations for the states in the state space formulation of the hydrological systems. For state X_t with input U_t, at time t, where the system is described by the function $f(\cdot)$ and the vector $\boldsymbol{\theta}$ of the (unknown) parameters, each state variable in the system equation is written

$$dX_t = f(X_t, U_t, t; \theta)dt + \text{“Noise”}dt, \tag{2.17}$$

where the noise term needs to be properly specified to obtain a sufficient uncertainty measures for the state. This equation is frequently employed in the succeeding chapters, as well as in most of the papers included.

CHAPTER 3

Grey box modelling

Over the last several decades, a variety of methods has been proposed for hydrological modelling, with the shared objective to predict the future response of the hydrological system in connection with, e.g., planning, designing or management (see an overview by *Singh and Woolhiser*, 2002). However, for both white box models and black box models the main concern is the uncertainty in the model structure and, as pointed out by *Refsgaard et al.* (2006), one of the main challenges for the future modelling aspects is to incorporate an adequate description of the uncertainty into the modelling framework. Hence, it is desirable to obtain a modelling approach that bridges the gap between the white box models in continuous-time and black box models in discrete-time, since this will facilitate the use of data in the modelling and subsequently in forecasting and control.

3.1 Modelling by stochastic differential equations

A grey box model is an approach that incorporates both white box and black box components. The fundamental equation in the grey box modelling approach is the Stochastic Differential Equation (SDE).

3.1.1 Stochastic differential equations

As observed in Chapter 2, the system dynamics of a hydrological process can often be interpreted by Ordinary Differential Equations (ODEs). In general the ODE is written

$$\frac{dX_t}{dt} = f(X_t, t; \boldsymbol{\theta}), \qquad t \geq 0 \tag{3.1}$$

and describes the dynamics of the variable X_t in a very rigid and deterministic fashion. A first attempt to add a stochastic part to the ODE (3.1) is simply by adding noise, and obtain

$$\frac{dX_t}{dt} = f(X_t, t; \boldsymbol{\theta}) + \sigma(X_t, t; \boldsymbol{\theta}) W_t, \quad t \geq 0 \tag{3.2}$$

where the additional term is a product of a given function $\sigma(X_t, t; \boldsymbol{\theta})$ and a reasonable stochastic process $\{W_t\}_{t \geq 0}$. A straightforward approach for the stochastic process would be to adopt stationary process with the properties $\mathrm{E}\{W_t\} = 0$, and W_t and W_s being independent for $t \neq s$. However, such a stochastic process cannot have continuous paths. A more appropriate approach is obtained by subdividing the time interval $[0, t]$:

$$0 = t_0 < t_1 < \ldots < t_k < \ldots < t_K = t, \tag{3.3}$$

and, subsequently, where $X_k = X_{t_k}$, $W_k = W_{t_k}$ and $\Delta t_k = t_{k+1} - t_k$, Eq. (3.2) can be rewritten on a discrete form as

$$X_{k+1} = X_k + f(X_k, t_k; \boldsymbol{\theta}) \Delta t_k + \sigma(X_k, t_k; \boldsymbol{\theta}) W_k \Delta t_k \tag{3.4}$$

with $k = 0, \ldots, K-1$. Replacing $W_k \Delta t_k$ with the stochastic term $\Delta \omega_t = \omega_{t+1} - \omega_k$ a suitable stochastic process $\{\omega_t\}_{t \geq 0}$ is obtained. This stochastic process should have stationary independent increments with mean zero, but the only suitable stochastic process with continuous paths that fulfills these requirements is the Wiener process (*Knight*, 1981).

The Wiener process is a fundamental continuous-time stochastic process for providing a mathematical interpretation of the diffusion processes. It is named after the American mathematician Norbert Wiener, but the process is often referred to as Brownian motion, in honour of Robert Brown[1]. Wiener's contribution to the mathematical theory of the Brownian motion led to the one-dimensional Brownian motion being referred to as a Wiener process. The main properties of the Wiener process are illustrated below, but more detailed descriptions of the process can be found by, e.g., *Madsen* (2008) and *Maybeck* (1982). The mathematical properties of the Wiener process, $\{\omega_t\}_{t \geq 0}$, are

[1] Brown was a botanist, who in 1828 was the first person to observe the irregular state of motion when he investigated the diffusion of small pollen grains

1. $P(\omega_0 = 0) = 1$
2. The (consecutive) increments of the process, for any partioning of the interval $0 \leq t_0 < t_1 < \ldots < t_N < \infty$, are mutually independent.
3. The Wiener process is Gaussian, i.e. the increments $\omega_t - \omega_s$ for any $0 \leq s < t$ is Gaussian with mean and covariance, respectively,

$$\begin{aligned} \mathrm{E}\{\omega_t - \omega_s\} &= 0 \\ \mathrm{V}\{\omega_t - \omega_s\} &= \sigma^2 |t - s| \end{aligned} \tag{3.5}$$

   where σ^2 is the incremental variance. The standard Wiener process is defined by $\sigma^2 = 1$.

The sample paths of the process are continuous with probability one, but are nowhere differentiable, also with probability one.

Now, by rewriting the discretised version (3.4) where the terms of the the standard Wiener process are included, the results for the whole process can be obtained by

$$X_K = X_0 + \sum_{k=0}^{K-1} f(X_k, t_k; \boldsymbol{\theta}) \Delta t_k + \sum_{k=0}^{K-1} \sigma(X_k, t_k; \boldsymbol{\theta}) \Delta \omega_k. \tag{3.6}$$

Letting $\Delta t_k \to 0$ and applying the traditional integration, given that the limit of the right hand side exists, the SDE can be solved by the integral

$$X_t = X_0 + \int_0^t f(X_s, s; \boldsymbol{\theta}) ds + \int_0^t \sigma(X_s, s; \boldsymbol{\theta}) d\omega_s \tag{3.7}$$

if – and only if – an appropriate interpretation of the second integral is provided.

The solution to (3.7) is found by integration, where the first integral can be defined in the traditional Riemann-Stieltjes sense. For the second integral, however, the variations of the paths of ω_t are too big for the integral to be properly defined in the Riemann-Stieltjes sense, i.e., the total variation of the path is almost surely infinite. To obtain a solution for the stochastic integral

$$\int_0^t \sigma(X_s, s; \boldsymbol{\theta}) d\omega_s$$

the sum of the Wiener processes is considered, i.e.

$$\sum_{k=0}^{K-1} \varphi(X_{\tau_k}, \tau_k; \boldsymbol{\theta}) \omega_{[t_k, t_{k+1})}(t) \tag{3.8}$$

where $\tau_k \in [t_k, t_{k+1})$ and the time discretisation is in accordance with (3.3) where $K \to \infty$. This sum cannot be interpreted in the Riemann-Stieltjes sense because of the unbounded variances of the Wiener paths. The approximation for the sum of Wiener processes (3.8) depends on the choice of the point τ_j, and selecting the left end point leads to the Itô integral where $\tau_k = t_k$ and the limit for the sum can be written

$$\sum_k \varphi\big(X_{t_k}, t_k; \boldsymbol{\theta}\big)\omega_{[t_k, t_{k+1})}(t) \to \int_0^t \sigma\big(X_\tau, \tau; \boldsymbol{\theta}\big) d\omega_\tau. \tag{3.9}$$

For modelling purposes with real data where SDEs are applied, the Itô integral is the most logical selection for interpretation of the stochastic integral, since it utilises present values in the integration (*Øksendal*, 2007). Alternative choice would be the Stratonovich integral where $\tau_k = (t_k + t_{k+1})/2$ in (3.8), which provides the integral with some nice properties from traditional calculus (*Stratonovich*, 1966, *Kloeden and Platen*, 1999). However, since the Stratonovich integral utilises future time-steps in the integral, it is not as sufficient for real-world applications. My entire thesis is based on real data, i.e., the focus is on solving (3.7) for real-world case studies. Therefore, only the Itô integral (3.9) is considered in the following.

For the stochastic variable X_t, the Itô process of the form in (3.7) is written on a differential form as a SDE:

$$dX_t = f(X_t, t; \boldsymbol{\theta})dt + \sigma(X_t, t; \boldsymbol{\theta})d\omega_t. \tag{3.10}$$

This is the form for the SDEs applied in the following case studies as an essential equation to describe the dynamics of the system. The system can then be described by several SDEs, where each SDE represents a state description in the model structure. The SDEs in Papers A-C, E and F apply (3.10) to formulate the state variables in the system descriptions. The first term on the right hand side of the SDE is the drift term. It describes the main part of the physical structure of the system and corresponds to the ODE in (3.1). The physical characteristics of the drift term are expressions most engineers are familiar with from formulating the traditional ODE models. The second term is the diffusion term of the SDE. It provides a suitable interpretation of the unavoidable dynamical errors, due to the fact that the mathematical model is not describing the true process in an exact way and inputs (or forcing) are also not known exactly. A large diffusion term in the SDE is usually caused by model approximations and partly known inputs that should be accounted for in the SDE.

3.1.2 Transforming the SDE

In the SDE in (3.10) the diffusion term is presented with a state dependency. Assessing the uncertainty of a system described by SDEs, indicates that the

uncertainty has to be accounted for in the SDEs. However, for physically-based systems the SDEs usually lead to a physical interpretation, which has to obey the physical laws. Thus, the diffusion term in the SDE has to correlate with the physics in the drift term in such a way that the uncertainty of the SDE is within the limitations of the response. An example of a state dependent diffusion is the sewer flow model in Papers E and F where the state variables in the model correspond to the volume of water stored in the reservoirs. This obviously has the lower limit zero since the volume cannot accept negative values. In other words, the diffusion of states must approach zero as the drift approaches the lower boundary of zero.

When the parameters are estimated in a model, where the system is described by SDEs with state dependent diffusion, some computational limitations prevent the estimation. For the estimation of SDEs with state dependent diffusion terms, higher order filtering techniques are required (*Vesteraard*, 1998). For the estimation, the software CTSM[2] is used (*Kristensen and Madsen*, 2003, *Kristensen et al.*, 2004a), but it applies an ordinary Kalman filter to evaluate the likelihood function for linear grey box models and an extended Kalman filter to obtain a solution for the nonlinear models. If a transformation is available for the SDE with a state dependent diffusion term, such that the diffusion becomes independent of the state, the filtering techniques in CTSM can be applied in order to obtain efficient and numerically stable estimates (*Baadsgaard et al.*, 1997).

The transformed SDE can be derived from a reasonable transformation of the stochastic variable X_t. Let $\phi(X_t)$ be a twice continuously differentiable function with respect to $X_t \in \mathbb{R}$ where $t \in \mathbb{R}_0$. Then a new stochastic variable $Z_t \in \mathbb{R}$ is defined as

$$Z_t = \phi(X_t)$$

where the SDE can be derived with a second order Taylor expansion of the Itô process of the variable X_t in (3.10). Hence, the SDE for Z_t is obtained by

$$dZ_t = \frac{\partial \phi(X_t)}{\partial t} dt + \frac{\partial \phi(X_t)}{\partial x} dX_t + \frac{1}{2} \frac{\partial^2 \phi(X_t)}{\partial x^2} (dX_t)^2. \tag{3.11}$$

By including the Itô process (3.10) in Itô's formula (3.11), and apply the rules

$$dt \cdot dt = dt \cdot d\omega_t = 0 \quad \text{and} \quad d\omega_t \cdot d\omega_t = dt,$$

the Itô process for the transformed variable Z_t can be written

$$\begin{aligned} dZ_t = &\left(\frac{\partial \phi(X_t)}{\partial t} + f(X_t, t; \boldsymbol{\theta}) \frac{\partial \phi(X_t)}{\partial x} + \frac{\sigma^2(X_t, t; \boldsymbol{\theta})}{2} \frac{\partial^2 \phi(X_t)}{\partial x^2} \right) dt \\ &+ \frac{\partial \phi(X_t)}{\partial x} \sigma(X_t, t; \boldsymbol{\theta}) d\omega_t. \end{aligned} \tag{3.12}$$

[2]Continuous-Time Stochastic Modelling - www.imm.dtu.dk/ctsm

To obtain a state independent diffusion term in the transformed Itô process the product in the diffusion of the transformed SDE (3.12) has to be equal to a term that is independent of X_t. However, such a term can still be considered as a function of the time and the parameters $\boldsymbol{\theta}$ in the original SDE (3.10), and is defined by $\widetilde{\sigma}(t;\boldsymbol{\theta})$. Then the aim is to find the diffusion term $\widetilde{\sigma}(t;\boldsymbol{\theta})d\omega_t$, but from (3.12) this corresponds to

$$\widetilde{\sigma}(t;\boldsymbol{\theta}) = \frac{\partial \phi(X_t)}{\partial x}\sigma(X_t,t;\boldsymbol{\theta}). \tag{3.13}$$

Rearranging this expression:

$$\frac{\partial \phi(X_t)}{\partial x} = \frac{\widetilde{\sigma}(t;\boldsymbol{\theta})}{\sigma(X_t,t;\boldsymbol{\theta})} \tag{3.14}$$

and to obtain the transformed variable Z_t that eliminates the random variable X_t from the diffusion in the original Itô process, Eq. (3.14) is integrated:

$$Z_t = \phi(X_t) = \widetilde{\sigma}(t;\boldsymbol{\theta}) \int \frac{dx}{\sigma(x,t;\boldsymbol{\theta})}\bigg|_{x=X_t}. \tag{3.15}$$

This transformation of a state with a univariate state dependency in the expression for the diffusion is referred to as the Lamperti transform (*Iacus*, 2008). By defining

$$\widetilde{f}(Z_t,t,\boldsymbol{\theta}) = \frac{\partial \phi(X_t)}{\partial t} + f(X_t,t;\boldsymbol{\theta})\frac{\partial \phi(X_t)}{\partial x} + \frac{\sigma^2(X_t,t;\boldsymbol{\theta})}{2}\frac{\partial^2 \phi(X_t)}{\partial x^2} \tag{3.16}$$

the transformed Itô process can be written

$$dZ_t = \widetilde{f}(Z_t,t,\boldsymbol{\theta})dt + \widetilde{\sigma}(t;\boldsymbol{\theta})d\omega_t. \tag{3.17}$$

The Lamperti transformation usually has the consequence that a SDE with a linear drift term, where the ordinary Kalman filter applies, becomes nonlinear, and an extended Kalman filter is required. However, by transforming the SDE, the unknown parameters in the original SDE are retained and can be estimated efficiently.

3.1.3 Stochastic grey box models

Formulation of a system that varies in both time and space and describes the movement of a physical phenomenon, is typically presented in the form of a partial differential equation (PDE). However, due to external factors that have substantial effect on the physical system, it is normally not possible to solve

the PDE analytically. Therefore, a simplification of the system formulation is required. One simplified approach is the lumped parameter model, obtained by replacing the PDE with a finite set of ODEs (as introduced by Eq. (3.1)), which can then by related to discrete time measurements by using a state space formulation:

$$d\boldsymbol{X}_t = \boldsymbol{f}(\boldsymbol{X}_t, \boldsymbol{U}_t, t, \boldsymbol{\theta})dt \tag{3.18}$$

$$\boldsymbol{Y}_k = \boldsymbol{g}(\boldsymbol{X}_k, \boldsymbol{U}_k, t_k, \boldsymbol{\theta}) + \boldsymbol{\varepsilon}_k, \tag{3.19}$$

where (3.18) is the system equation, describing the variation in time of the physical state in the system in continuous-time, and (3.19) is the observation equation that indirectly relates the observations to the states in discrete-time. The time $t \in \mathbb{R}_O$ indicates the continuous time and k ($k = 1, \ldots, K$) is the discretely observed sampling instants for K number of measurements. $\boldsymbol{U} \in \mathbb{R}^m$ is a vector of input variables and $\boldsymbol{Y} \in \mathbb{R}^l$ is a vector of the output variables. The state variables $\boldsymbol{X} \in \mathbb{R}^n$ represent the dynamic behaviour of the system, where the system dynamics are determined by the function $\boldsymbol{f}(\cdot) \in \mathbb{R}^n$. The function $\boldsymbol{g}(\cdot) \in \mathbb{R}^l$ in the observation equation describes how the output variables are functions of the indirectly observed state variables $\boldsymbol{X}_k$ with residuals $\boldsymbol{\varepsilon}_k$. The vector $\boldsymbol{\theta} \in \mathbb{R}^p$ contains the unknown parameters in the system.

Presenting the system equation as a finite set of ODEs indicates that the uncertainty is not accounted for in the system structure, and the residuals $\boldsymbol{\varepsilon}_k$ contain not only the measurement noise, but also noise terms related to model approximation in the system description, undetected input variables and insufficient input measurements. Consequently, an autocorrelation is detected in the sequence of the output noise terms $\boldsymbol{\varepsilon}_k$. To improve the existing model, a separation has to be made between the noise terms assigned to the model and the input approximations (referred to as process noise), and the errors directly related to the observations. Replacing the set of ODEs in the system equation with a set of SDEs provides this separation of the model output noise into a process noise, which is now accounted for in the diffusion terms of the system equation, and a measurement noise. Thus, the grey box model is introduced:

$$d\boldsymbol{X}_t = \boldsymbol{f}(\boldsymbol{X}_t, \boldsymbol{U}_t, t, \boldsymbol{\theta})dt + \boldsymbol{\sigma}(\boldsymbol{X}_t, \boldsymbol{U}_t, t, \boldsymbol{\theta})d\boldsymbol{\omega}_t \tag{3.20}$$

$$\boldsymbol{Y}_k = \boldsymbol{g}(\boldsymbol{X}_k, \boldsymbol{U}_k, t_k, \boldsymbol{\theta}) + \boldsymbol{e}_k, \tag{3.21}$$

where the entry of the process noise is described by $\boldsymbol{\sigma}(\cdot) \in \mathbb{R}^{n \times n}$; the measurement error $\boldsymbol{e}_k$ is assumed to be a l-dimensional white noise process with $\boldsymbol{e}_k \in N(\boldsymbol{0}, \boldsymbol{V}(\boldsymbol{U}_k, t_k, \boldsymbol{\theta}))$ and $\{\boldsymbol{\omega}_t\}$ is a n-dimensional standard Wiener process. Hence, the grey box model provides the necessary separation between the process noise and the measurement noise. This noise separation is illustrated in Figure 3.1. Similar to (3.19), the observation equation in the grey box model (3.21) relates the discrete time observations to the state variables at times where observations are available. When determining unknown parameters of the

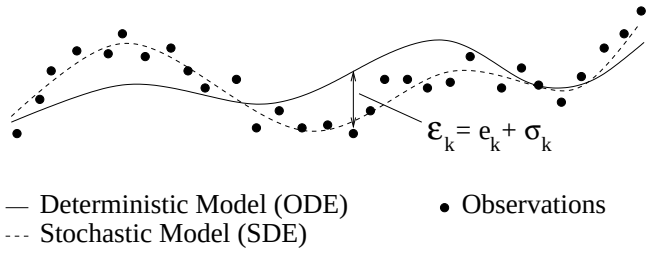

Figure 3.1: Partitioning of the output noise ε_k, into the process noise σ_k and the measurement noise e_k. Replacing the ODEs in the system equation by SDEs provides an improved model description.

model from a set of data, the continuous time formulation provides the model with flexibility consisting of possibilities for varying sample times and for missing observations in the data series.

One form of the grey box model in (3.20) and (3.21) is to consider the functions $\boldsymbol{f}$ and $\boldsymbol{g}$ linear, i.e.

$$\boldsymbol{f}(\boldsymbol{X}_t, \boldsymbol{U}_t, t; \boldsymbol{\theta}) = \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{X}_t + \boldsymbol{B}(\boldsymbol{\theta})\boldsymbol{U}_t \tag{3.22}$$

$$\boldsymbol{g}(\boldsymbol{X}_t, \boldsymbol{U}_t, t; \boldsymbol{\theta}) = \boldsymbol{C}\boldsymbol{X}_t \tag{3.23}$$

where the matrices $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ describe the dynamics of the linear system. Furthermore, the diffusion term is simplified, such that it is only a function of the unknown parameters in the model. Then the grey box model becomes linear and time-invariant:

$$d\boldsymbol{X}_t = [\boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{X}_t + \boldsymbol{B}(\boldsymbol{\theta})\boldsymbol{U}_t]\,dt + \boldsymbol{\sigma}(\boldsymbol{\theta})d\boldsymbol{\omega}_t \tag{3.24}$$

$$\boldsymbol{Y}_k = \boldsymbol{C}(\boldsymbol{\theta})\boldsymbol{X}_k + \boldsymbol{e}_k. \tag{3.25}$$

The system equation (3.24) corresponds to the stochastic groundwater model introduced in Eq. (2.16) as a set of SDEs with a Wiener process to represent the additional stochastic noise in the system.

This grey box modelling approach is used in Paper A and Paper B. These papers have the joint objective to introduce the grey box modelling approach as a tool to analyse and model flows in well fields. Here, the focus is on the drift term in the system equation, and an effort is made to illustrate the advantage of applying statistical tools to identify and locate the flaws in a lumped parameter model structure for a single well in the well field. As an example: in Paper A the resulting residual series for the model output is autocorrelated, as Figure 3.2 reveals a significant autocorrelation on the first time-lag, and this implies that an additional state is needed to obtain an adequate model structure to describe the water head in the well. Applying this, the simple model

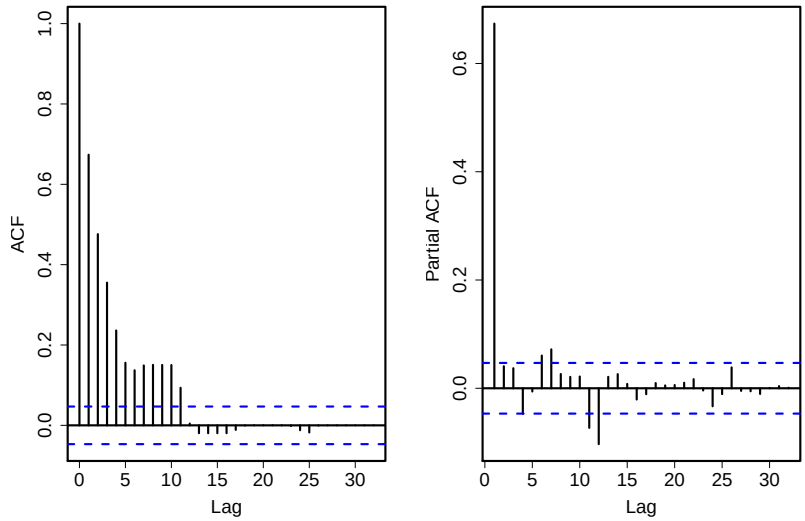

Figure 3.2: The autocorrelation and partial autocorrelation for the residuals from modelling the water head in the well of interest.

can be extended towards an expression of the spatio-temporal variation of the groundwater flow in the whole well field. However, no further model extensions were made by this introduced step-by-step procedure to include more details of the physical meaning of the system, since this methodology has been thoroughly presented (see e.g. *Kristensen et al.*, 2004b,a, *Jonsdottir et al.*, 2006b). Also, for large spatial models based on the grey box approach, a class which uses the mathematical algorithm behind CTSM is required.

Instead of including additional states into the simple lumped parameter model for the well field, a little different approach is shown in Paper C, where the lumped parameters in an existing drift term are investigated and expanded in order to improve the physical structure of the model. Furthermore, the aim in Paper C is also to provide the system equation a suitable description of the uncertainty, for which is obtained by assuming the diffusion term to be time-variant. The states are not considered to be directly observed and the observation equation depends on the input variables. Thus, the linear modelling framework in Eq's. (3.24) and (3.25) is extended, and becomes

$$d\boldsymbol{X}_t = [\boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{X}_t + \boldsymbol{B}(\boldsymbol{\theta})\boldsymbol{U}_t]\,dt + \boldsymbol{\sigma}(\boldsymbol{U}_t, t, \boldsymbol{\theta})d\boldsymbol{\omega} \tag{3.26}$$

$$\boldsymbol{Y}_k = \boldsymbol{C}(\boldsymbol{\theta})\boldsymbol{X}_k + \boldsymbol{D}(\boldsymbol{\theta})\boldsymbol{U}_k + \boldsymbol{e}_k. \tag{3.27}$$

where the matrix $\boldsymbol{D}(\boldsymbol{\theta})$ relates the measured input to the output variables. The challenge is to come up with a suitable description for the diffusion in the

model, which can be applied to reduce the prediction intervals of the water levels as time moves further away from the time of decision for the pumping rate. Thus, the diffusion in the system is expressed by an exponential function, which is initiated every time a pumping rate is changed for any of the wells in the well field. This approach provides the model with a measure for the uncertainty that increases with the decisions taken regarding ability of the pumps to meet the water demand, but does decrease fairly rapidly when the time between decisions prolongs.

The effect of improving the existing model structure by extending the lumped parameters in the simple model, and of including the exponential function in the diffusion, respectively, is plotted in Figure 3.3. The figure shows very clearly how the prediction interval for the models is improved as the model is developing from the lumped parameter model to a more detailed parameter model where the diffusion term is given a proper function in order to cope with the uncertainty in the model structure.

For the modelling approach of the sewer flow in Paper E and Paper F, the drift term is assumed to be linear and time-invariant, but the diffusion term in the system equation has a state-dependency. Hence, the model is written as in the general case in Eq's. (3.20) and (3.21), where the functions $\boldsymbol{f}$ and $\boldsymbol{g}$ are written as in (3.22) and (3.23), respectively. However, for the software CTSM, used for the parameter and state estimation of the grey box models in order to obtain feasible estimates, it is a necessary condition that the diffusion $\boldsymbol{\sigma}$ is independent of the states $\boldsymbol{X}_t$. Thus, to estimate the model parameters, the grey box model is transformed in order to remove the state dependencies from the diffusion terms in the system equation.

In the sewer runoff model the diffusion for the individual state variable is considered to be a function of only the state itself, which indicates that the Lamperti transform can be used to obtain a state independent system equation (see Section 3.1.2). Thus, the transformed grey-box model, applied in Paper E and Paper F, becomes

$$d\boldsymbol{Z}_t = \widetilde{\boldsymbol{f}}\big(\boldsymbol{Z}_t, \boldsymbol{U}_t, t; \boldsymbol{\theta}\big)dt + \widetilde{\boldsymbol{\sigma}}\big(\boldsymbol{U}_t, t, \boldsymbol{\theta}\big)d\boldsymbol{\omega}_t \tag{3.28}$$

$$\boldsymbol{Y}_k = \widetilde{\boldsymbol{g}}(\boldsymbol{Z}_k, \boldsymbol{U}_k, t_k; \boldsymbol{\theta}) + \boldsymbol{e}_k \tag{3.29}$$

where $\boldsymbol{Z}_t$ is a vector including the transformed states at time t and the function $\widetilde{\boldsymbol{f}}$ is now a nonlinear description for the drift terms of the transformed state space model; $\widetilde{\boldsymbol{g}}$ is describing the observation equation, but now as a function of the transformed states, and $\widetilde{\boldsymbol{\sigma}}$ is a state independent diffusion term. The system description consists of two linear reservoirs in a series, such as introduced in (2.8). A state dependent diffusion is included in the system equation for the

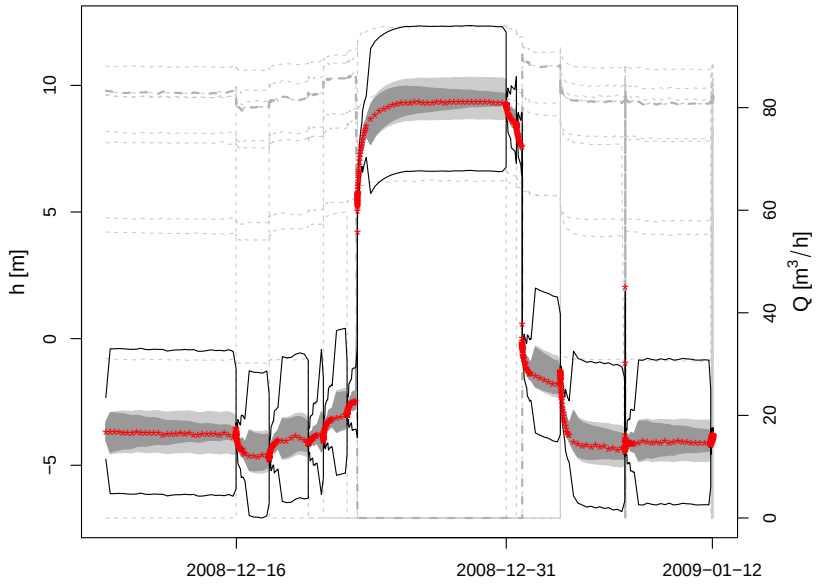

Figure 3.3: Comparison between prediction interval of the three models in Paper C. The interval for the lumped model is between the two black lines, the interval for the model where the drift term has been improved is the light grey area, and the model with the extended diffusion term is illustrated with the dark grey area. The observations are marked with red stars, and the pumping rate for all the wells (with the corresponding well highlighted) are the dashed lines.

sewer runoff and can be written

$$d\boldsymbol{X}_t = \boldsymbol{f}(\boldsymbol{X}_t, \boldsymbol{U}_t, t, \boldsymbol{\theta})dt + \begin{bmatrix} \sigma_1 X_{1,t}^{\gamma_1} & 0 \\ 0 & \sigma_2 X_{2,t}^{\gamma_2} \end{bmatrix} d\boldsymbol{\omega}_t. \tag{3.30}$$

To transform the system equation so that the diffusion term becomes state independent, the Lamperti transform

$$Z_{i,t} = \frac{X_{i,t}^{1-\gamma_i}}{1-\gamma_i} \qquad i = 1,2$$

is applied and subsequently Itô's formula (3.11) provides the transformed states. In this study, three models are proposed, which are different in the diffusion term in the system equation (3.30). The diffusion only deviates in the γ parameters: Model 1 has a constant diffusion ($\gamma_i = 0$, for $i = 1,2$, in (3.30)); Model

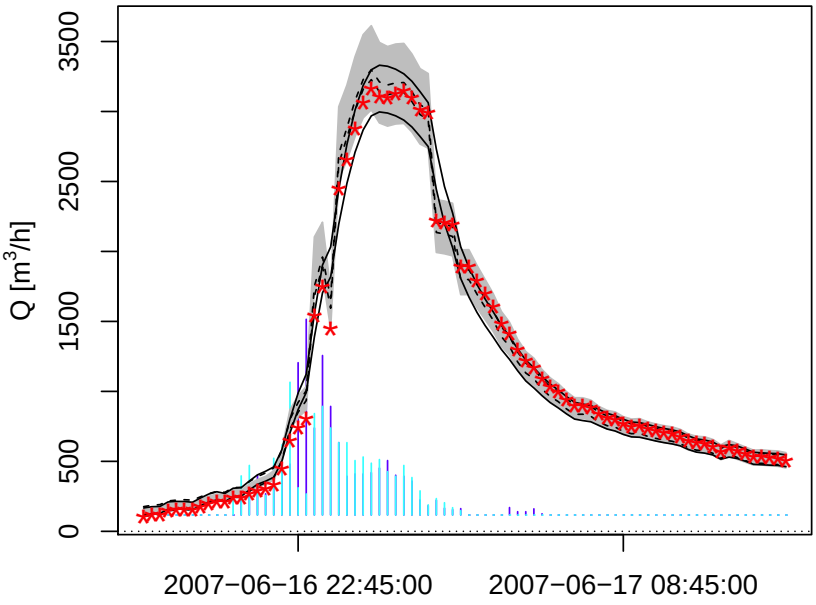

Figure 3.4: The prediction intervals for the three proposed models in Paper E. The dash lines correspond to the interval of a model with a constant diffusion term; the solid lines represent the limits of a model with state dependent diffusion, with $\gamma \in (0.5, 1)$; and the grey area is the prediction interval where $\gamma_1 = \gamma_2 = 1$. The measured flow is displayed with red stars, and the magnitude of the measured rain from the two closest gauges are the blue and cyan coloured barplots.

2 has a proportional state dependency (γ_1 and γ_2 equal 1); and Model 3 has a value between 0.5 and 1. The model uncertainty, included in the diffusion, is detected in the output uncertainty. This uncertainty for a single rain event in the time series is illustrated in Figure 3.4 and it clearly shows the improved performance obtained by extending the diffusion in the system equation.

3.2 Impulse response function models

In the thesis, an alternative modelling approach is applied to model the water levels in the wells in the well field. For modelling the water levels as a direct function of the pumping rates, an Impulse Response Function (IRF) model is

considered, but an IRF model is very different from the grey box model introduced in the previous section. The IRF has been mentioned in Section 2.3.1 where hydrographs are described, but the unit hydrograph is defined as the IRF of a catchment outlet flow when a unit impulse of rain enters the watershed. Similar phenomenon occurs in the well field, when a particular well is observed at the same time as the pumping rate is changed in another operating well penetrating the same aquifer. In the observed well the water level responds, but the time delay of the response (the retention time) is according to the physical characteristics of the aquifer and the distance between the two wells.

An IRF is a non-parametric description of the linear system. For a linear and time-invariant system, the output $Y(t)$ at time $t \geq 0$ can be obtained in continuous time by the convolution integral

$$Y(t) = \int_{-\infty}^{\infty} \theta(\tau) U(t-\tau) d\tau + N(t) \tag{3.31}$$

where $U(t-k)$ is the measured system input at time $t-k$, $k \leq t$, and $N(t)$ is a correlated noise term. Here, $\theta(k)$ is a weight function that represents the IRF, considered at the lagged time k.

In Paper D IRFs are used to model the piezometric head in the wells, with well i being influenced by all pumping wells in the well field. Since water is pumped from a confined aquifer the model can be considered linear. For a given series of pumping for the N wells in the well field, the responding water level in well i becomes

$$h_i(t) = \sum_{j=1}^{N} \int_{-\infty}^{t} Q_j(\tau) \theta_{ij}(t-\tau) d\tau + b_i(t) + N_i(t), \tag{3.32}$$

where θ_{ij} is the IRF of well i depending on the pumping rate in well j, and $b_i(t)$ is considered as an upper boundary of the water head in well i and corresponds to a scenario with no pumping in any of the wells. In continuous time, the impulse response for the water drawdown is detected as the changes in $h_i(t)$ for a unit impulse of discharge Q_j, and since an increase in pumping rate results in a decrease in the water head, the IRF can be defined as

$$\theta_{ij} = -\frac{\partial h_i}{\partial Q_j} \qquad i,j = 1,\ldots,N. \tag{3.33}$$

Discretising the convolution integral (3.31) offers an approach for determining the shape of the IRF, as expressed by a rational polynomials (*Box and Jenkins*, 1970, *Madsen*, 2008). However, keeping the model in continuous time enables a parameterisation that, to some extent, has a reasonable physical meaning, and

by applying so-called PIRFICT models[3] (*von Asmuth et al.*, 2002) the IRFs are defined as simple parametric analytical expressions. One approach in continuous time is defined by the Hantush formula, describing the penetration in well i in an aquifer of infinite extent that responds to an operating well j at distance r_{ij}. The IRF is formulated as

$$\theta_{ij}(t) = \frac{1}{4\pi Tt} \exp\left(-\frac{r_{ij}^2 S}{4Tt} - \frac{t}{cS}\right) \tag{3.34}$$

where the parameters are transmissivity T [L^2T] and storage coefficient S [-], and a parameter of a storage-free aquitard with resistance c [T] covering the penetrated aquifer.

For the shape of the IRF $\theta_{ij}(t)$, the IRF in (3.34) is simplified with regard to the physical parameters and an alternative expression is proposed. Since the shape of the function should only depend on the measured distance r_{ij}, the IRF becomes

$$\theta_i(t) = -\frac{A}{t^\beta} \exp\left(-\frac{\beta\lambda_i}{t}\right), \tag{3.35}$$

derived from (3.34) where A and β are constant for each well. The characteristics of this expression for the IRF are well suited for modelling the water level in the wells, e.g., for all wells $\theta_i(0) = 0$ with first order derivative; global optimum is reached at $t = \lambda_i$, corresponding to the peak flow (or peak delay) in the hydrograph (Figure 2.3); and for $t \to \infty$, $-A/t^\beta \to 0$ and the IRF asymptotically decays to zero. The IRFs are illustrated graphically in Figure 3.5, but the difference between the impulse response of two pumping wells i and j is determined by the difference between the peak flows λ_i and λ_j.

The Hantush equation is also used to stimulate a description of the diffusion term in Paper C. This is considered feasible since the equation provides an exponential decay to reduce the estimated prediction intervals, and also because it is expressed by the parameters already included in the grey box model.

3.3 Maximum likelihood estimation

The Maximum Likelihood (ML) method is a very flexible and efficient statistical method for estimating unknown parameters in a model. Given the sequence of the measured output

$$\mathcal{Y}_K = \left[\boldsymbol{Y}_K, \cdots, \boldsymbol{Y}_k, \cdots, \boldsymbol{Y}_1, \boldsymbol{Y}_0\right] \tag{3.36}$$

[3] Predefined Impulse Response Function In Continuous Time

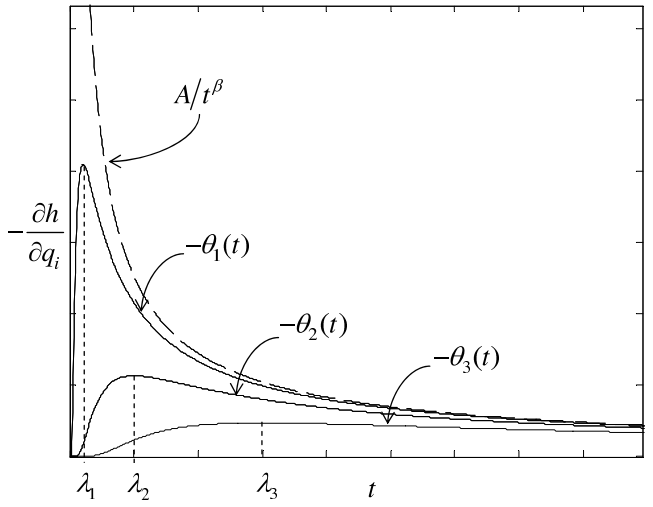

Figure 3.5: The role of parameters A, β and λ_i in shaping the IRF of equation (3.35).

ML estimates of the unknown parameters can be determined by finding the set of parameters $\boldsymbol{\theta}$ that maximises the likelihood function

$$L(\boldsymbol{\theta};\mathcal{Y}) = p(\mathcal{Y}|\boldsymbol{\theta}), \tag{3.37}$$

i.e. the conditional probability of obtaining the observed sequence given the parameter set $\boldsymbol{\theta}$ (*Madsen and Thyregod*, 2011). This indicates that the likelihood function is simply the joint probability distribution function for all observations, which for time series data can be written

$$L(\boldsymbol{\theta};\mathcal{Y}_k) = \left(\prod_{s=1}^{k} p(\boldsymbol{Y}_s|\mathcal{Y}_{s-1},\boldsymbol{\theta})\right) p(\boldsymbol{Y}_0|\boldsymbol{\theta}), \tag{3.38}$$

where $p(A,B) = p(A|B)p(B)$ is applied to express the likelihood as a product of conditional densities. The parameter estimates can be determined by conditioning on the initial values and solving the optimisation problem

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta}\{\ln\left(L(\boldsymbol{\theta};\mathcal{Y}_N|\boldsymbol{Y}_0)\right)\}. \tag{3.39}$$

Most often it is not possible to optimise the likelihood function analytically, and hence numerical methods have to be applied to obtain the optimal parameter set for the maximised likelihood function (*Madsen*, 2008).

To estimate the parameters in the stochastic grey box model in (3.20) and (3.21) a filtering method is applied, which seeks to approximate solutions to the

continuous-discrete time nonlinear filtering problem. Since the diffusion term is assumed to be independent of the state variables, the Extended Kalman Filter (EKF) can be applied. The Gaussian density is completely characterised by its mean and covariance denoted by

$$\hat{\boldsymbol{Y}}_{k|k-1} = E\{\boldsymbol{Y}_k|\mathcal{Y}_{k-1},\boldsymbol{\theta}\}$$

and

$$\boldsymbol{R}_{k|k-1} = V\{\boldsymbol{Y}_k|\mathcal{Y}_{k-1},\boldsymbol{\theta}\},$$

respectively, and by introducing an expression for the innovation:

$$\boldsymbol{\epsilon}_k = \boldsymbol{y}_k - \hat{\boldsymbol{y}}_{k|k-1}$$

the likelihood function can be rewritten as

$$L(\boldsymbol{\theta};\mathcal{Y}_N) = \left(\prod_{k=1}^{N} \frac{\exp\left(-\frac{1}{2}\boldsymbol{\epsilon}_k^\top \boldsymbol{R}_{k|k-1}^{-1}\boldsymbol{\epsilon}_k\right)}{\sqrt{\det(\boldsymbol{R}_{k|k-1})}\left(\sqrt{2\pi}\right)^l} \right) p(\boldsymbol{Y}_0|\boldsymbol{\theta}) \tag{3.40}$$

where the conditional mean and covariance are calculated using either a Kalman Filter for linear models or an Extended Kalman Filter for nonlinear models. The likelihood (3.40) is used in Papers A, B, C, E and F in order to find the optimal parameter set for the proposed grey box models.

By using the maximum likelihood method to estimate the parameters in the grey box model in (3.20) and (3.21) is a rather straight forward procedure. However, it is not easily solved due to the numerical optimization that searches for the solution in multidimensional parameter space (*Kristensen et al.*, 2004b). To solve the estimation problem the previously mentioned open source software CTSM is used.

For the parameter set in the IRF models (here also referred to as $\boldsymbol{\theta}$) in Paper D the estimation scheme is a little different, since the objective of the IRF model is to produce scenarios for decisions to be applied to the controlled pumping rates, where the time between decisions is Δt. This indicates that the time between every two decisions has to be simulated, and included in the parameter estimation. For the continuous-time IRF models the noise series for the output error can also be considered in continuous-time, given by the stochastic convolution integral (*von Asmuth et al.*, 2002):

$$N(t) = \int_{-\infty}^{t} \phi(t-\tau)d\omega(\tau)$$

where $\omega(t)$ is the Wiener process (see Section 3.1.1) and $\phi(t)$ is an exponential IRF. The sequence of innovations $\{\epsilon(t)\}_{t\geq 0}$ is obtained from the simulated

output error sequence, i.e.

$$\begin{aligned}\epsilon(t) &= \int_{\Delta t}^{t} \phi(t-\tau) d\omega(\tau) \\ &= N(t_i|t_1) - e^{-\alpha\Delta t} N(t_{i-1}|t_1)\end{aligned}$$

and the conditional likelihood function can be found by a similar expression as in (3.38), or

$$L(\boldsymbol{\theta}; \mathcal{Y}_K) = \frac{1}{(2\pi\sigma_\epsilon^2)^{\frac{K}{2}}} \prod_{k=1}^{K} \exp\left(\frac{-\epsilon^2(t_k)}{2\sigma_\epsilon^2}\right). \tag{3.41}$$

with σ_ϵ^2 as the variance of the innovation series.

3.3.1 Uncertainty in parameter estimates

The maximum likelihood method provides an assessment of the uncertainty for the parameter estimates, attained from the fact that by the central limit theorem the estimator in (3.39) is asymptotically normal distributed with mean $\boldsymbol{\theta}$ and covariance matrix

$$\hat{\Sigma}_{\boldsymbol{\theta}} = \boldsymbol{H}^{-1}.$$

The matrix $\boldsymbol{H}$ is the information matrix, given by

$$h_{ij} = -E\left\{ \frac{\partial^2}{\partial\theta_i \partial\theta_j} \ln\left(L(\boldsymbol{\theta}|\mathcal{Y}_{k-1})\right) \right\} \qquad i,j = 1,\ldots,p. \tag{3.42}$$

Due to the asymptotic Gaussianity of the estimator in (3.39) a t-test can be performed for significance of the estimated parameters. Subsequently, the likelihood function can be tested statistically in search for the most appropriate model by using a likelihood ratio test (*Madsen and Thyregod*, 2011). For a proposed model, the test can be used to determine if the model performs significantly better than a more simple model, where the parameter space for the simple model is a subgroup of the one for the proposed model. A sequence of such a likelihood ratio tests for model selection provides a stopping criterion for the model development, resulting in a model that renders the best fit to data (*Bacher and Madsen*, 2011).

3.4 Discussion

One of the main advantages of the grey box approach lies in the SDEs for the model formulation. A classical hydrological model on state space form is represented by a set of ODEs, where the only noise term for the whole model is

detected in the observation equation. By replacing the ODEs with SDEs, a noise term is assigned to each state in the system description that quantifies the lack of fit for each state of the model. Thus, the presented grey box model introduces a system description that is physically meaningful and the parameters contain physical interpretation. Also, the approach facilitates that the parameter estimates are calibrated based on the input-output measurements.

The procedure contains statistical tools to verify the model. This indicates that data for both input and output is required for estimating the model parameters, as if the model structure was purely stochastic. Compared with the white box models, the necessity of parameter calibration can be omitted if the model is defined within a well-established hydrological system. However, the parameter adjustment in the grey box model is crucial due to the fact that all mathematical models are approximations of the true process, where a part of the lack in fit is due to assumptions in the hydrological parameters of the system. A further argument is uncertainty in measurements of the input, or forcing of the system. This is well known for hydrological models (see e.g. *Beven*, 1989, *Harremoës and Madsen*, 1999, *Radwan et al.*, 2004, *Refsgaard et al.*, 2006), but with the grey box model approach the statistical methods can be observed as verification tools for both the model and the parameters in the system, i.e., the significance of the parameters and the identifiability of the model is in accordance with the input-output relation in the available measurements. Accordingly the statistical methods are utilised for model validation, and a subsequent model selection if that is required (*Kristensen et al.*, 2004a, *Møller et al.*, 2010a, *Bacher and Madsen*, 2011).

The problem of identifying the model structure errors, and errors due to forcing of the system has been pointed out as one of the most difficult challenges in deriving at conceptual models that adequately can predict the outcome of the hydrological system (*Refsgaard and Henriksen*, 2004, *Refsgaard et al.*, 2010). Moreover, if the model structure error is identifiable, and has significant influence on the model output, measures have to be considered in order to deal with the error in such a way that it is not detected in the model output. The grey box model approach provides a partitioning of the prediction error into error terms, which are directly related to the model structure and the forcing of the system, and an output error term. Also, by representing the conceptual model on a state space form, as a set of SDEs, the model structure error and the forcing uncertainty is further separated and assigned to the states in such a way that the error term for each state variable can be quantified and, subsequently formulated to fulfill the physical requirements for the hydrological system.

CHAPTER 4

Prediction, uncertainty and evaluation

The key to successful predictions of hydrological events is a decent underlying stochastic model. Proposing the grey box model for forecasting supplies short-term predictions with a sufficient description of the uncertainty. Furthermore, the included significant physical knowledge will sustain the long-term effects as the prediction horizon is extended. Thus, the focus in this chapter is on predictions based on grey box models that trigger the importance of assessing the uncertainties due to the approximations in the model structure and the errors in the input (forcing) specification.

4.1 Predictions using grey box models

With a proposed stochastic model the objective is to predict the output at time $k+h$. The measured output at time $k+h$ is denoted as Y_{k+h}. In parallel, we have $\hat{Y}_{k+h|k}$ as the prediction of the output at time $k+h$, given the available information at time k where h indicates the horizon for the prediction. With the given sequence of input up to time $k+h$ as $\mathcal{U}_k = [\boldsymbol{U}_k, \cdots, \boldsymbol{U}_0]^\top$ and the output sequence up to time k as $\mathcal{Y}_k = [\boldsymbol{Y}_k, \ldots, \boldsymbol{Y}_0]^\top$, the optimal prediction in a least square sense is equal to the conditional mean (see proof by *Madsen*, 2008).

Hence, the prediction is obtained by

$$\begin{aligned}\hat{\boldsymbol{Y}}_{k+h|k} &= E\{\boldsymbol{Y}_{k+h}|\mathcal{Y}_k,\mathcal{U}_{k+h}\}\\ &= \boldsymbol{g}\left(\mathcal{Y}_k,\mathcal{U}_{k+h},t_{k+h},\boldsymbol{\theta}\right).\end{aligned}$$

If the input is not known for future values a modification is needed (see again *Madsen*, 2008).

Concerning the grey box model described in Eq's. (3.20) and (3.21): Due to the indirectly observed states in the system equation, the observation equation is a function of the state variables. Hence, to predict the model outcome, predictions for the states have to be provided and included in the observation equation:

$$\hat{\boldsymbol{Y}}_{k+h|k} = \boldsymbol{g}\left(\hat{\boldsymbol{X}}_{k+h|k},\mathcal{U}_{k+h},t_{k+h},\boldsymbol{\theta}\right), \tag{4.1}$$

where $\hat{\boldsymbol{X}}_{k+h|k}$ is the state predictor. Thus, the challenge in predicting the future outcome in the system is not directly related to predictions using the observation equation, but instead to predictions for the state variables in the system equation. To predict the states only inputs from k to $k+h$ are required. The state prediction can be accomplished by considering the conditional expectation of the future state, i.e.,

$$\hat{\boldsymbol{X}}_{k+h|k} = E\{\boldsymbol{X}_{k+h}|\hat{\boldsymbol{X}}_{k|k},\boldsymbol{U}_{k+h},\cdots,\boldsymbol{U}_k\}, \tag{4.2}$$

where $\hat{\boldsymbol{X}}_{k|k}$ is the reconstruction of the state at time k, given all measurements to the same time k (*Madsen*, 2008). For states that are observable the measured values are used instead of the reconstruction, i.e. $\boldsymbol{X}_k = \hat{\boldsymbol{X}}_{k|k}$ in (4.2). The state reconstruction is also referred to as filtering of the unmeasured states in the model, providing a mean and a variance for the normally distributed state at each time instant k. For predictions at time $k+h$ the distributional properties of the state variable at time k must to be taken into account.

Obtaining a sufficient input sequence for the predictions h time steps ahead can be problematic. However, this is not an issue for predictions of well water levels in a well field (Papers C and D) because the input is a sequence of decisions for the operating wells, and is determined beforehand. Issues related to the input are raised for predictions of sewer runoff where the flow is influenced by rain (Papers E and F), but weather conditions are fairly unpredictable phenomena. Today the most popular measurement devices for rainfall are rain gauges. The rain gauge provides a time series of rainfall up to time k and one possibility is to construct a time series model to forecast the input h steps ahead, depending on past and present values of rain. However, such a model can easily become rather complex, since the spatio-temporal variation of rain is

highly dependent on external variables, e.g. meteorological variables such as wind speed and wind direction. Models for the predicted input is not a part of my thesis and, thus, in Paper E and especially in Paper F, where the prediction horizons are 4 hours ahead, the rain is assumed to be known a priori, and is used as an input to the model to generate the predictions.

4.1.1 Numerical solution of SDEs for predictions

The normal assumption for the model output is only valid for one-step ahead predictions. Thus, for $h > 1$ a numerical approach is considered where an Euler scheme is applied for the SDEs in the system equation (3.20) in order to simulate predictions for the states in the system (*Kloeden and Platen*, 1999):

$$\hat{\boldsymbol{X}}_{k+\Delta|k} = \hat{\boldsymbol{X}}_{k|k} + \boldsymbol{f}\left(\hat{\boldsymbol{X}}_{k|k}, \boldsymbol{U}_{k+\Delta}, \boldsymbol{\theta}\right)\Delta + \boldsymbol{\sigma}\left(\hat{\boldsymbol{X}}_{k|k}, \boldsymbol{U}_{k+\Delta}, \boldsymbol{\theta}\right)\Delta W_k. \tag{4.3}$$

Δ is the time step for the Euler approximation, and ΔW_k is a randomly generated increment of the Wiener process $\{W\}_k$, i.e. $\Delta W_k = W_{k+\Delta} - W_k$. The Euler scheme in (4.3) is presented for simulations of one-step ahead predictions, but to obtain an Euler scheme for h-step ahead, (4.3) is extended and written

$$\begin{aligned}\hat{\boldsymbol{X}}_{k+h|k} =& \hat{\boldsymbol{X}}_{k|k} + \left(\sum_{i=1}^{h/\Delta} \boldsymbol{f}\left(\hat{\boldsymbol{X}}_{k+(i-1)\Delta|k}, \boldsymbol{U}_{k+i\Delta}, \boldsymbol{\theta}\right)\right)\Delta \\ &+ \sum_{i=1}^{h/\Delta} \boldsymbol{\sigma}\left(\hat{\boldsymbol{X}}_{k+(i-1)\Delta|k}, \boldsymbol{U}_{k+i\Delta}, \boldsymbol{\theta}\right)\Delta W_{k+(i-1)\Delta}.\end{aligned} \tag{4.4}$$

With an increasing prediction horizon the variance of the stochastic term increases and the accuracy of a single point prediction, generated from (4.4), is reduced.

In Papers C and E only the one-step ahead prediction is considered for the well water level and the sewer runoff, respectively. The predictions are provided by using (4.3), whereas the four hour ahead predictions of the sewer runoff in Paper F utilises the extended Euler scheme (4.4).

4.2 Prediction intervals

It is common procedure for both deterministic and stochastic models to quantify prediction performances by only considering the distance between the forecasts and corresponding observations, often referred to as point prediction performance (*Madsen et al.*, 2005). However, the information obtained from the

point predictions is not sufficient to capture the information embedded in the considered prediction interval. Thus, to obtain a probability distribution for the h-step ahead prediction, Eq. (4.4) is used to generate a number of simulations to profile a predictive distribution. The number of simulations has to be large so that the predictive distributions can generate reasonable prediction intervals for the state variables. For the predicted output in the observation equation to be adequately evaluated, a proper assessment for the prediction uncertainty is imposed in order to cope with the varying state variables in the model, i.e., a reasonable prediction interval for the model output is obtained from the simulated prediction intervals for the states in the system equation. The prediction interval for the output Y_{k+h} is then defined as quantiles of the simulated outcomes.

In the following, the ideal coverage of the prediction interval is defined as the nominal coverage $1-\beta$, $\beta \in [0,1]$. The upper and lower limits of the interval prediction are obtained from quantile forecasts determined on the basis of the large number of simulations for the state predictors. This results in an empirical probability distribution for the model output. If $F_{k+h|k}$ is the cumulative distribution function of the predicted output $\hat{Y}_{k+h|k}$, and $\tau \in [0,1]$ is the proportion of the relative quantile, the quantile forecast for the $k+h$ prediction is obtained by

$$q^{(\tau)}_{k+h|k} = F^{-1}_{k+h|k}(\tau). \tag{4.5}$$

It is required that the prediction intervals are properly centered on the probability density function. Usually, the median to the predictive distribution is chosen, implying that there is equal probability for each simulation, generated for the lead time $k+h$, to be below or above the estimated intervals, i.e., $\beta/2$ is left outside the coverage on each side of the prediction interval. This is well suited for shorter prediction horizons, but for longer horizons this needs to be assessed by studying the predictive distribution of the forecast quantiles. If $l = \beta/2$ and $u = 1-\beta/2$ are defined as the lower and upper quantiles for the prediction interval at level $1-\beta$, respectively, then the prediction interval for the lead time $k+h$, issued at time k, can be described as

$$\hat{I}^{(\beta)}_{k+h|k} = \left[\hat{q}^{(l)}_{k+h|k}, \hat{q}^{(u)}_{k+h|k}\right] \tag{4.6}$$

where $\hat{q}^{(l)}_{k+h|k}$ and $\hat{q}^{(u)}_{k+h|k}$ are, respectively, the lower and upper prediction limits at levels $\beta/2$ and $1-\beta/2$ (*Pinson et al.*, 2007, *Møller et al.*, 2008).

4.3 Evaluation of prediction intervals

To evaluate the performance of the grey box model it is important to include the assessed prediction interval of the model output in the evaluation criterion. The most common evaluation criteria are, e.g., (root) mean square error, mean average error (*Madsen et al.*, 2005) and in hydrology the Nash-Sutcliffe coefficient (*Nash and Sutcliffe*, 1970) is usually applied. However, none of these criteria can be applied to assess the uncertainty of the prediction since any quantification related to the prediction interval is omitted. Thus, to include the interval in the criterion; reliability, sharpness and resolution are introduced, but these three measures all influence the skill score criterion that is used for the evaluation. With such a measure for the performance of the predictive abilities of a specific model, different models can be compared and subsequently decisions can be made that are not only based on the prediction of the model, but also the interval characteristics.

Reliability, sharpness and resolution have been addressed before, both directly and indirectly, in connection with evaluation of hydrological models. This has mostly been done in relation to modelling uncertainties with the Generalized Likelihood Uncertainty Estimation (GLUE) method (*Beven and Binley*, 1992); a method that has been applied for a variety of environmental systems. In general, reliability is a measure of bias between the model and the measurements, since it quantifies the percentages of measurements within a given quantile. In the following, the reliability is quantified as the proportion of observations within given coverage, and corresponds to the containing ratio introduced by *Xiong et al.* (2009). *Jin et al.* (2010) also used the reliability and resolution concepts as adopted in the following, but additionally they propose the Average Relative Interval Length (ARIL) as a measure for the concentration of the prediction interval. The difference between the sharpness and ARIL is that the ARIL is inverse proportional to the measured flow and, thus, a qualitative measure of the relative sharpness, whereas the sharpness is a quantitative measure. However, the aim here is to obtain a quantitative measure that can be applied to interpret the interval skill score criterion, which then corresponds to an overall evaluation for the prediction interval.

4.3.1 Reliability

In order for the prediction interval to be of any practical usage for decision makers, it is a primary requirement that the interval is reliable to such an extent that the upper and lower limits have to correspond to the nominal coverage rate of $1 - \beta$. To obtain an evaluation of the reliability of the interval, a counter is defined that rewards prediction intervals capable of capturing the

observations. For a given prediction interval, as formulated in (4.6), and a corresponding measured output y_{k+h}, an indicator variable is obtained by

$$n_{k,h}^{(\beta)} = \begin{cases} 1, & \text{if } y_{k+h} \in \hat{I}_{k+h|k}^{(\beta)} \text{ for } k \leq K-h \\ 0, & \text{otherwise.} \end{cases} \tag{4.7}$$

Considering all observations a binary time series $\{n_{k,h}^{(\beta)}\}$ is obtained, corresponding to hits and misses of the prediction interval. The mean of the binary series then represents the actual proportion of hits in the whole time series. For prediction horizon h the proportion of hits, for a series of length K, is given by

$$\bar{n}_h^{(\beta)} = \mathrm{E}\left[n_{k,h}^{(\beta)}\right] = \frac{1}{K-h}\sum_{k=1}^{K-h} n_{k,h}^{(\beta)}. \tag{4.8}$$

The accuracy between the nominal coverage and the proportion of hits is defined as the reliability of the prediction interval, denoted by

$$b_h^{(\beta)} = 1 - \beta - \bar{n}_h^{(\beta)}, \tag{4.9}$$

where the perfect fit is defined as $b_h^{(\beta)} = 0$ and the reliability is fulfilled. When discrepancy is detected between the empirical coverage and the theoretical one, the coverage of the prediction interval is biased, for which $\bar{n}_h^{(\beta)} > 1-\beta$ is considered as an overestimated bias in the coverage and $\bar{n}_h^{(\beta)} < 1-\beta$ an underestimated bias.

The reliability is illustrated in Figure 4.1, which displays the reliability of the five different models that are analysed in Paper F. The reliability is plotted as a function of the coverage and clearly shows that for all models the bias is increased towards overestimation when the coverage is increased to around 40%. For further increasing coverages the bias is reduced, but for 90-95% coverage most of the models are considered to be underestimated. To be accurate, only two models fulfill the reliability; Model 3 for 80% coverage and Model 5 for 90% coverage. However, for the 85-90% coverage all models can be defined within the range of the reliability and, therefore, these coverages can be considered optimal.

4.3.2 Sharpness and resolution

Sharpness is a measure of the accuracy of the prediction interval where smaller values indicate that the model is better suited to generate predictions (*Gneiting et al.*, 2007). As sharpness approaches zero, more weight is put on the accuracy

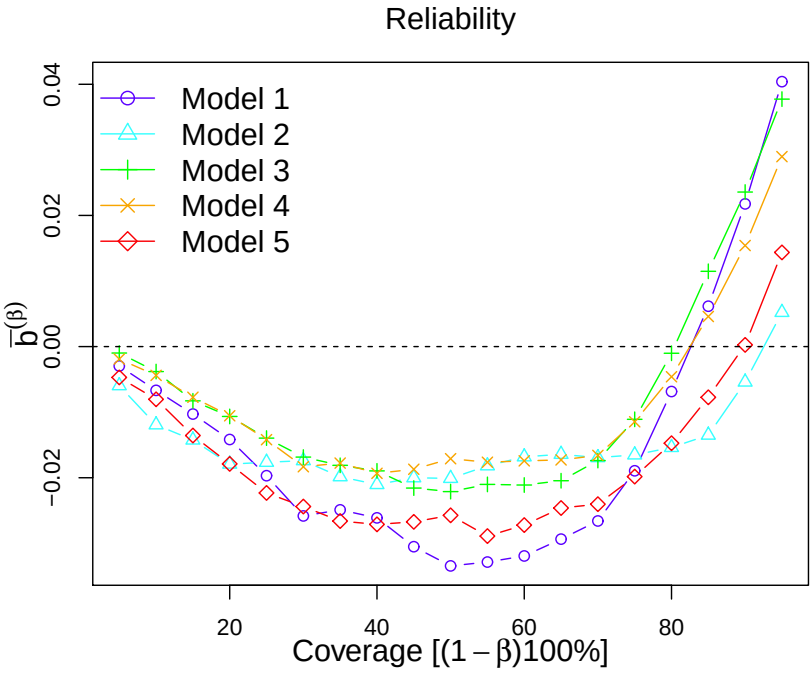

Figure 4.1: Reliability of the five models in Paper F, plotted as a function of the coverage.

of using point predictions. Thus, the size of the interval predictions serves as a measure of sharpness of the predictive distribution. The size of the interval prediction, issued at time k for lead time $k+h$, is defined as the difference between the corresponding upper and lower quantile forecast, and averaging for the whole time series defines the sharpness. For the h horizon and coverage $1-\beta$, the sharpness is calculated by

$$\overline{\delta}_h^{(\beta)} = \frac{1}{K}\sum_{k=1}^{K}\left(\hat{q}_{k+h|k}^{(u)} - \hat{q}_{k+h|k}^{(l)}\right) \tag{4.10}$$

and by calculating $\overline{\delta}_h^{(\beta)}$ for relevant coverages, a δ-diagram can be viewed to summarise the evaluation of the sharpness.

Resolution is defined as the potential for obtaining different predictive distributions if dependence on forecast conditions is taken into account. Since it is required that the predictive distribution must be reliable, it can be further stated that the resolution is characterised by its ability to provide distinct predictive distributions, depending on the conditional reliability.

For the drainage flows in Paper F the input variables are rain, observed by rain gauges. Thus, the most obvious forecast condition is to distinguish between wet- and dry-weather situations. For evaluation, δ-diagrams (and also reliability) for different groupings of the forecast conditions can be drawn for comparison for the average shape of the predictive distribution. δ-diagrams for both the unconditional sharpness and the conditional sharpness for dry and wet weather situations, are displayed in Figure 4.2. When the prediction horizon increases, it is expected that the sharpness will increase as well, since the distribution of the prediction is expanding with the horizon. Without conditioning, the least sharpest prediction is obtained by using Model 2, whereas the sharpest ones are brought on by Model 1 and Model 3. If the sharpness is conditioned on the presence of rainfall events, the sharpness is drastically increased as shift occurs from dry to wet weather situations (lower panels in Figure 4.2). In wet weather the sharpness is at least twice the size of the sharpness in dry weather, where the shift has different influence on the models. The sharpness in dry weather is similar to sharpness without conditioning, because the number of time instants in the dry weather is 90% of the entire time series. Thus, by only considering the unconditional sharpness (and reliability) the important performance measures for the rain events in the time series are not detected. In wet weather, Model 1 is now the only model with the sharpest prediction, whereas Model 5 is approaching Model 2 both being the model with the greatest prediction interval.

Based on the measure of sharpness alone, it is difficult to reach conclusions regarding the prediction performances. Coming up with an appropriate prediction interval is not a straightforward approach since too narrow limits leave out too many observations, and the reliability of the prediction is lost. With too wide intervals the predictions become infeasible basis for decision making. Since the sharpness is exclusively a property of the prediction interval, no information is provided for the predictive distribution, as compared to the observations as these become available at time $k + h$. As the observations become available, they cannot be disregarded in the performance evaluation of the prediction intervals.

4.3.3 Unique skill score

To obtain a quantitative measure for the performance of the models, a scoring criterion is required that takes into account the prediction and corresponding observation and discrepancies between the model and the measurements. An appropriate criterion would be the skill score criteria, which gathers all the information of the proposed model into a single numerical value for the model performance (*Gneiting and Raftery*, 2007). Such a performance measure would

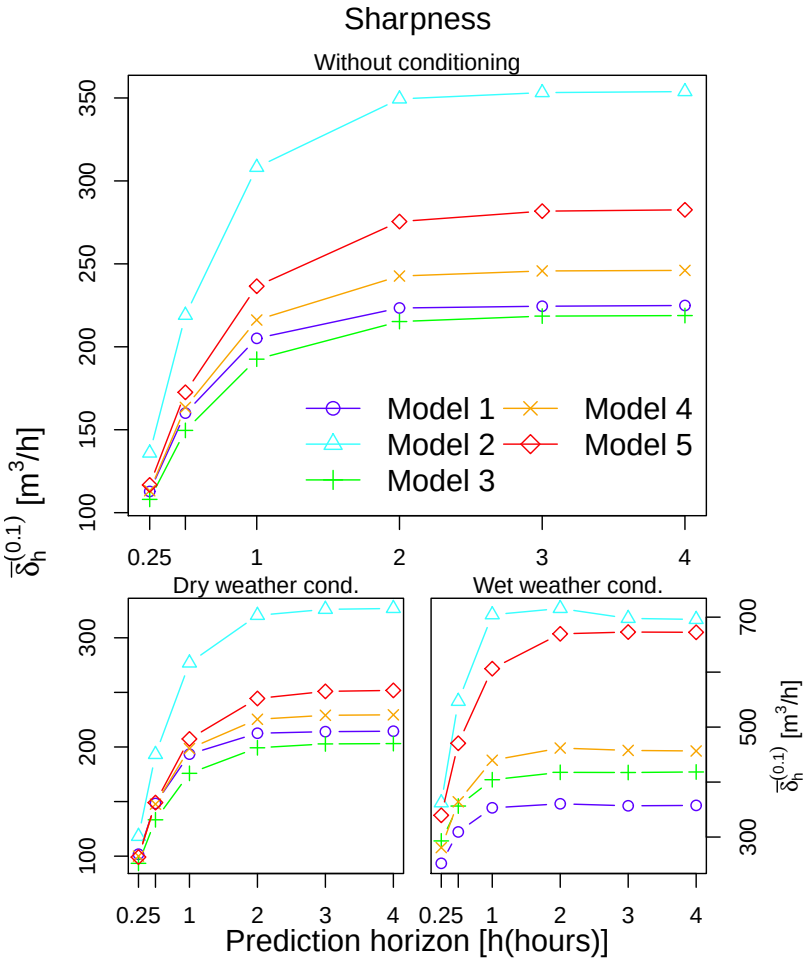

Figure 4.2: The sharpness is plotted as a function of the increasing prediction horizon – both independent of and conditioned on the weather situations for the five models in Paper F. Top: The sharpness of the predictions, without distinguishing between dry and wet weather conditions; bottom left: Sharpness for the dry-weather flow; bottom right: Sharpness for the wet-weather flow (rain events).

be given by the score $Sc(Q,Y)$ to a predictive distribution Q if the event Y materialises. The expected score under the probability measure P is defined as

$$Sc(Q;P) = \int Sc(Q;Y)dP(Y) \tag{4.11}$$

for Q observed and opposed to the predictive distribution P. A scoring rule is said to be proper if a prediction corresponds to the forecaster's judgement. This indicates that if the aim is to minimise the skill score over a validation set, the score is proper if for any two distributions Q and Q'

$$Sc(Q,Q') \geq Sc(Q,Q), \qquad \forall Q,Q'.$$

If the equal sign is included in the equation, the scoring rule is said to be strictly proper.

If probabilistic forecasts are represented by quantile forecasts – by considering the quantiles $q_1,\cdots,q_l$ for the proportions $\tau_1,\cdots,\tau_l$, respectively – the skill score criterion in (4.11) can be written

$$Sc(q_1,\cdots,q_l;P) = \int Sc(q_1,\cdots,q_l;Y)dP(Y) \tag{4.12}$$

and is referred to as quantile skill score. From the quantile skill score the interval skill score is obtained by only considering the set of quantiles that form the interval in (4.6) in the skill score in (4.12). The skill score Sc for the interval prediction, at time instant k, is calculated as (*Gneiting and Raftery*, 2007)

$$\begin{aligned} Sc^{(\beta)}_{I,k,h} =& sc(\hat{I}^{(\beta)}_{k+h|k};Y_{k+h}) = (\hat{q}^{(u)}_{k+h|k} - \hat{q}^{(l)}_{k+h|k}) \\ &+ \frac{2}{\beta}(\hat{q}^{(l)}_{k+h|k} - Y_{k+h})\mathbb{1}\{Y_{k+h} < \hat{q}^{(l)}_{k+h|k}\} \\ &+ \frac{2}{\beta}(Y_{k+h} - \hat{q}^{(u)}_{k+h|k})\mathbb{1}\{Y_{k+h} > \hat{q}^{(u)}_{k+h|k}\}. \end{aligned} \tag{4.13}$$

where the indicator $\mathbb{1}(\cdot)$ equals one if the included statement holds. Otherwise, it is zero. It can be seen from (4.13) that the skill score is increased for any observation that is outside the predefined prediction interval. Thus, the skill score gives a positive penalisation, and with model comparison the best performing model is the one with the lowest skill score.

The score criterion in (4.13) regards only a single time step in the output series, but since the objective is to evaluate the prediction in total by a single number, an extension is required in order to account for the entire series. Usually, this is done by aggregating the scores for all time instants where observations are available, either by summation or averaging. The benefits of summing up for the entire time series are apparent when performance of different models is compared where the score values become more distinct. This is the case for the model comparison in Paper C where the results become more decisive by summing up all score values.

However, considering the average score of the time series the following applies: Firstly, the average interval score can be directly described by the sharpness and, secondly, the average interval score becomes independent of the

length of the time series. The average interval score criterion is written

$$\begin{aligned}\bar{Sc}_{I,h}^{(\beta)} =& \frac{1}{K}\sum_{k=1}^{K} Sc_{I,h,k}^{(\beta)} = \bar{\delta}_h^{(\beta)} \\ &+ \frac{2}{\beta(K-h)}\sum_{k=1}^{K-h}\Big[(\hat{q}_{k+h|k}^{(l)} - Y_{k+h})\mathbb{1}\{Y_{k+h} < \hat{q}_{k+h|k}^{(l)}\} \\ &+ (Y_{k+h} - \hat{q}_{k+h|k}^{(u)})\mathbb{1}\{Y_{k+h} > \hat{q}_{k+h|k}^{(u)}\}\Big].\end{aligned} \tag{4.14}$$

The second term in this scoring criterion shows that the score is increased for an observation outside the predicted interval in the magnitude of the distance between the interval and observation. The indication of the individual observation in relation to the interval can be merged into an indicator, corresponding to the reliability indicator (4.7). Thus, the interval score (4.14) can be written as an indirect function of the prediction interval in (4.6) by including the indicator, i.e.,

$$\begin{aligned}\bar{Sc}_{I,h}^{(\beta)} =& \bar{\delta}_h^{(\beta)} + \frac{2}{\beta(K-h)}\sum_{k=1}^{K-h}\left(1 - n_{k,h}^{(\beta)}\right) \\ &\times \left(\min\left|Y_{k+h} - [\hat{q}_{k+h|k}^{(l)}, \hat{q}_{k+h|k}^{(u)}]\right|\right)\end{aligned} \tag{4.15}$$

where the second term under the summation accounts for the minimum distance between the observed value and the prediction interval, which is always the lower or the upper limit of the interval.

The score criterion in (4.15) is used in Paper F, both for comparing the performances of the models and to compare the performances of the wet-weather flow and the dry-weather flow. Even though many more observations were detected in the dry-weather flow periods, the interval skill scores for the two different flow regimes can be directly compared.

The score is still a function of the prediction horizon h. This indicates that there are just as many $\bar{Sc}_{I,h}^{(\beta)}$ as there are h. To evaluate the performance, independent of h, we simply average all horizons, thus obtaining the interval score criterion $\overline{\bar{Sc}}_I^{(\beta)}$.

To demonstrate the interval skill score, I continue with the evaluation from the study in Paper F. Table 4.1 displays the score values for the prediction horizons and the five model proposals, but comparison with the sharpness in Figure 4.2 shows the importance of quantifying the deviation of the individual target miss from the estimated prediction interval. This can best be detected by observing Model 1 and 3. These models generated the sharpest prediction intervals, but

the calculated skill score is rather poor. Even though Model 3 is the best fit for the first two lead times, the increased score values for the larger prediction horizons resulted in an overall skill score that is significantly higher than the one of the optimal prediction model, namely Model 5. Hence, on average, and for almost all horizons, Model 5 has the lowest skill score and is the favoured model candidate for the predictions.

4.4 Discussion and conclusions

By using the included model approaches, this chapter has given a general overview over the methods used for providing predictions. The focus is on the interval predictions to adequately assess the uncertainty. In particular, prediction intervals for the grey box model is dealt with since the uncertainty of both the system description and the forcing of the system can be embedded in the system structure by considering the separation between model and input approximations, and the measurement noise. The improved description for the uncertainty in the model structure is the key element to obtain sufficient predictive distributions for outputs from hydrological systems, since the uncertainty varies with the prediction horizon. Hence, simulations using grey box models provide probabilistic forecasts for future scenarios, such that reasonable prediction intervals can be attained.

A well determined evaluation criterion for the prediction abilities has to account for the prediction intervals along with the point prediction for the outcome. However, the predictive ability of the model is comprised in the model structure. For the evaluation, reliability, sharpness and resolution are tools that the modeller is provided so that he or she is able to adequately signify the lack of fit in the model predictions. The evaluation of the models in Paper F clearly demonstrates this property of the evaluation measures. The resolved

Table 4.1: Skill score values for the 90% prediction intervals in Paper F. The best score value for each horizon is highlighted.

	Prediction Horizon						
	0.25h	0.5h	1h	2h	3h	4h	Average
Model 1	166.0	292.7	491.2	680.8	724.7	732.7	514.7
Model 2	201.9	324.9	455.2	563.6	602.6	610.1	459.7
Model 3	**137.2**	**228.3**	391.2	603.8	675.8	691.7	454.7
Model 4	155.1	264.3	429.7	606.4	663.6	673.6	465.4
Model 5	150.4	247.1	**383.8**	**535.2**	**593.8**	**608.2**	**419.7**

evaluation results in a large bias and sharpness for the wet weather conditions, indicating that the simple linear reservoir model is too simple to describe the flow from rain to corresponding runoff in the drainage system. These findings can then be verified by comparing the skill score values for the wet and dry weather conditions, but the introduced average skill score in (4.15) can be used directly to compare the differently resolve time series for the same model. For a model to be equally advantageous for prediction for all groups of conditioning, the skill score values should be close to one another.

CHAPTER 5

Conclusions and further perspectives

The objective of the thesis was formulation of stochastic dynamic hydrological systems, where the focus was on modelling and interpreting the uncertainties embedded in the model structure. Two modelling approaches were applied, a stochastic differential equation based model and an impulse response function model. The first approach was formulated as a continuous-discrete time stochastic state-space model, where the states were represented by stochastic differential equations, which consist of the drift term that corresponds to the ordinary differential equations describing the dynamics of the hydrological system, and the diffusion term accounting for the uncertainty in the time evolution of the states due to, e.g., model approximations and uncertainty in the measurements of the input (forcing). By formulating the diffusion terms, the uncertainty could be assigned to the related states and, thus, improve the prediction intervals of the model output. This approach is referred to as grey box modelling because it bridges the gap between the physically-based model in continuous time and the statistical model in discrete time. The impulse response function models were applied in continuous time, where the parameters were provided a physical interpretation by using known equations from hydrology as impulse response functions. Thus, applying a similar argument as the one applied to the previous method, the impulse response function model, as presented here, can in a way also be considered a grey box model. The unknown parameters in both methods are estimated by the maximum like-

lihood method. However, these two modelling approaches are rather different, since the states in the output from the stochastic state-space model are related to the input variables via the states in the model formulation. For the impulse response function models the output is a direct consequence of changes in the input sequence.

The performances of the stochastic differential equation based models were evaluated by three measures: reliability, sharpness and resolution. For the performance comparison, a more global approach was exploited; the skill score criterion, which gathers the properties of the prediction in a single number. The criterion accounts for the prediction interval, but not only the point prediction of the output, i.e., the assessed uncertainty is dealt with in the evaluation criterion. The skill score penalises predictions with too comprehensive estimates for the prediction interval. Also, a measurement that is detected outside the prediction interval is penalised proportionally in accordance with the deviation from the interval.

Two case studies were used, although within fairly contrasting hydrological fields: well field modelling and sewer runoff modelling. For the well field case study, both modelling methods were applied to a water head response in operating wells in a well field, where the pumping rates from the wells were the only available input series. Both approaches gave promising results for predicting the water head in the wells along with appropriate measures of uncertainties, despite the fact that the available data for modelling was rather limited and defected, and required substantial cleaning before it was approved for modelling purposes (*Dorini et al.*, 2011). The two model approaches have its advantages when it comes to groundwater management. The stochastic state-space approach can adequately describe the embedded uncertainty and is the optimal selection for forecasting and control of the well field. Due to the drift of the model being formulated from the main physical structure of the hydrological system, the long-term effects are also attained in relation to the details of the drift term. However, the model only includes wells that penetrate the same aquifer, and the parameter estimation of the stochastic state-space model is rather time consuming with the tools available today. On the other hand, for the impulse response function models the main advantage lie in simulations for the entire well field using a model that is robust and can generate results in a very short time. Uncertainties do not have the same constructive interpretation as is the case for the stochastic state-space formulation of the grey box model.

Considering the grey box model as a possibility for future modelling of groundwater management has several benefits, but further work is required to fully describe the properties of the well field. For the stochastic state-space model, the approach is limited to a single aquifer and, therefore, rather site-specific. A

natural next step would be to include interactions between the various aquifers in the well field that are penetrated and supply water to the water distribution network. The modelling approach would then take a step towards a more generalised framework for stochastic modelling of well fields. Consequently, the approach becomes feasible for real-time control of the water systems, where the uncertainty of the physical characteristics and the forcing of the system need to be comprehended and included in the model structure. Also, the available data for the case study in the thesis contained so-called on-off pumps, only, but these types of pumps are being replaced by frequency pumps that can be tuned to a preferred discharge rate. For future modelling of the stochastic well field model this has to be incorporated in the uncertainty of the model structure, where one option is to consider a state dependent diffusion in the state descriptions.

For the sewer runoff modelling, the stochastic differential equation based model was considered. The suggested diffusion terms in the model included a state dependency that was taken care of by Lamperti transforming the states to obtain state independent diffusion terms. This resulted in a model with varying prediction intervals to describe the increasing uncertainty of the flow as rainfall is detected in the catchment. Both the model and its parameters were physically interpretable and identifiable from data, but the evaluation measures for the model (reliability, sharpness and resolution) revealed a lack of fit for larger prediction horizons when rain events occurred. This is not surprising, since the rainfall-runoff process is a fairly complex hydrological phenomenon that is hardly totally obtainable with a single series of linear reservoirs, as well as the rain gauges available for the study are located outside the catchment, which causes an increasing uncertainty in the rain input. Thus, to obtain improved and more general results for the sewer runoff from the catchment, the next step would be to extend the drift term in the model structure, but the model extensions for the drift term also have to be accounted for in the diffusion term, since the extensions indicate that more physical constraints have to be fulfilled and reflected in the diffusion of the model structure. Furthermore, as the uncertainty of the flow is highly dependent on the rain events, it is also feasible to account for the rain input in the diffusion terms in the model structure.

To summarise, the simple grey box model, where the diffusion is given more attention is a reasonable approach to predict the outcome of the hydrological systems where measures of the embedded uncertainties, in form of model approximations and uncertainty in the system forcing, are required. The framework enables a stepwise procedure to detect the lack of fit in the grey box model, and combined with the uncertainty assessment introduced in this thesis, the grey box model approach provides more robust models that are better suited for forecasting and control in managing hydrological systems.

Bibliography

Adams RA (1999) *Calculus: a complete course*, 4th Edition. Addison Wesley Longman Ltd., Canada, Ch. 16, pp. 946–953.

Anderson MP, Woessner WW (2002) *Applied Groundwater Modeling - Simulation of Flow and Advective Transport*. Academic Press, San Diego, California, USA.

Baadsgaard M, Nielsen JN, Spliid H, Madsen H, Preisel M (1997) Estimation in stochastic differential equations with state dependent diffusion term. In: *SYSID '97 - 11th IFAC symposium of system identification, IFAC*.

Bacher P, Madsen H (2011) Identifying suitable models for the heat dynamics of buildings. *Energy and Buildings* **43** (7):1511–1522.

Beven K (1989) Changing ideas in hydrology - the case of physically-based models. *Journal of Hydrology* **105**:157–172.

Beven K, Binley AM (1992) The future of distributed models: model calibration and uncertainty prediction. *Hydrological Processes* **6**:279–298.

Box GEP, Jenkins GM (1970) *Time Series Analysis, Forecasting and Control*. Holden-Day, San Francisco.

Chow VT, Maidment DR, Mays LW (1988) *Applied Hydrology*. McGraw-Hill Book Company.

Dorini GF, Thordarson FÖ, Madsen H, Madsen H (2011) *Analysis and treatment of the søndersø time series - grey box well field modelling*. Tech. Rep. IMM-Technical Report-2011-04, Technical University of Denmark - DTU Informatics.

Douglas JF, Gasiorek JM, Swaffield JA (2001) *Fluid Mechanics*, 4th Edition. Prentice Hall - an imprint of Pearson Education, Essex, England.

Gneiting T, Balabdaoui F, Raftery AE (2007) Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B* **69** (2):243–268.

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association* **102** (477):359–378.

Gupta RS (2008) *Hydrology and Hydraulic Systems*, 3rd Edition. Waveland Press, USA.

Harremoës P, Madsen H (1999) Fiction and reality in the modelling world - balance between simplicity and complexity, calibration and identification, verification and falsification. *Water Science and Technology* **39**(9):1–8.

Iacus SM (2008) *Simulation and Inference for Stochastic Differential Equations - with R Examples*. Springer series of Statistics.

Jin X, Xu C-Y, Zhang Q, Singh VP (2010) Parameter and modeling uncertainty simulated by glue and a formal bayesian method for a conceptual hydrological model. *Journal of Hydrology* **383**:147–155.

Jonsdottir H, Madsen H, Eliasson J, Palsson OP (2006a) Conditional parametric models for storm sewer runoff. *Water Resources Research* **43**:1–9.

Jonsdottir H, Madsen H, Palsson OP (2006b) Parameter estimation in stochastic rainfall-runoff models. *Journal of Hydrology* **326** (1-4):379–393.

Kloeden PE, Platen E (1999) *Numerical Solution of Stochastic Differential Equations*. Springer-Verlag.

Knight FB (1981) *Essentials of brownian motion and diffusion*. American Mathematical Society.

Kristensen NR, Madsen H (2003) *Continuous Time Stochastic Modeling - CTSM 2.3 - Mathematics Guide*. Technical Unversity of Denmark.

Kristensen NR, Madsen H, Jørgensen SB (2004a) A method for systematic improvement of stochastic grey-box models. *Computers and Chemical Engineering* **28** (8):1431–1449.

Kristensen NR, Madsen H, Jørgensen SB (2004b) Parameter estimation in stochastic grey-box models. *Automatica* **40**:225–237.

Madsen H (2008) *Time Series Analysis*. Chapman & Hall/CRC.

Madsen H, Gudbjerg J, Falk AK (2008) A combined groundwater and pipe network model for well-field management. In: *MODFLOW and more: Ground water and Public Policy*. May 19-21, Golden, Colorado, USA.

Madsen H, Pinson P, Kariniotakis G, Nielsen HA, Nielsen TS (2005) Standardizing the performance evaluation of short-term wind power prediction models. *Wind Engineering* **29** (6):475–489.

Madsen H, Thyregod P (2011) *Introduction to General and Generalized Linear Models*. Chapman & Hall/CRC.

Maybeck P (1982) *Stochastic Model, Estimation and Control*. Vol. 1, 2, & 3. Academic Press, New York, USA.

McDonald MG, Harbaugh AW (1983) *A modular three-dimensional finite-difference ground-water flow model*. Tech. rep., U.S. Geological Survey.

Møller JK, Madsen H, Carstensen J (2010a) Structural identification and validation in stochastic differential equation based models - with application to a marine ecosystem np-model. Submitted.

Møller JK, Nielsen HA, Madsen H (2008) Time-adaptive quantile regression. *Computational Statistics & Data Analysis* **52**:1292–1303.

Møller JK, Phillipsen KR, Christiansen LE, Madsen H (2010b) Development of a resticted state space stochastic differential equation model for bacterial growth in rich media. Submitted.

Nash JE (1957) The form of the instantaneous unit hydrograph. *IASH* **3**:144–121.

Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models part i - a discussion of principles. *Journal of Hydrology 10* (3):282–290.

Øksendal B (2007) *Stochastic Differential Equations - An Introduction with Applications*, 6th Edition. Springer-Verlag.

Philipsen KR, Christiansen LE, Hansen H, Madsen H (2010) Modelling conjugation with stochastic differential equations. *Journal of Theoretical Biology* **263**:134–142.

Pinson P, Nielsen HA, Møller JK, Madsen H (2007) Non-parametric probabilistic forecasts of wind power: Required properties and evaluation. *Wind Energy* **10** (6):497–516.

Radwan M, Willems P, Berlamont J (2004) Sensitivity and uncertainty analysis for river quality modelling. *Journal of Hydroinformatics* **6** (2):83–99.

Refsgaard JC, Henriksen HJ (2004) Modelling guidelines - terminology and guiding principles. *Advances in Water Resources* **27**:71–82.

Refsgaard JC, Højberg AL, Møller I, Hansen M, Søndergaard, V. (2010) Groundwater modeling in integrated water resources management - visions for 2020. *Ground water* **48** (5):633–648.

Refsgaard JC, van der Sluijs JP, Brown J, van der Keur P (2006) A framework for dealing with uncertainty due to model structure error. *Advances in Water Resources* **29**:1586–1597.

Rozos E, Koutsoyiannis D (2010) Error analysis of a multi-cell groundwater model. *Journal of Hydrology* **392**:22–30.

Sherman LK (1932) Streamflow from rainfall by the unit-graph method. *Engineering News Record* **108**:501–505.

Singh VP, Woolhiser DA (2002) Mathematical modeling of watershed hydrology. *Journal of Hydrologic Engineering* **7** (4):270–292.

Søgaard HT (1993) *Stoachstic systems with embedded parameter variations - applications to district heating*. Ph.D. thesis, Technical Unversity of Denmark - DTU, IMSOR.

Stratonovich RL (1966) A new representation for stochastic integrals and equations. *J. Siam Control* **4**:362–371.

Vesteraard M (1998) *Nonlinear filtering in stochastic volatility models*. Master's thesis, Technical University of Denmark - DTU, Department of Mathematical Modelling, Lyngby, Denmark.

Viessman W, Lewis GL (1996) *Introduction to hydrology*. HarperCollins College Publishers, New York, USA.

von Asmuth JR, Bierkens MFP, Maas K (2002) Transfer function-noise modeling in continuous time using predefined impulse response functions. *Water Resources Research* **38** (12):23.1–23.12.

Xiong L, Wan M, Wei X, O'Connor KM (2009) Indices for assessing the prediction bounds of hydrological models and application by generalised likelihood uncertainty estimation. *Hydrological Sciences Journal* **54** (5):852–871.

Part II

Papers

PAPER A

Grey box modelling of a groundwater well field

Authors:
F. Ö. Thordarson, H. Madsen, H. Madsen

In preceedings:
ModelCare (2009)

Grey Box Modeling of a Groundwater Well Field

Fannar Örn Thordarson[1], Henrik Madsen[1], Henrik Madsen[2]

Abstract

A modelling framework, called grey box modelling, is presented, which combines prior physical knowledge of dynamic groundwater well field systems with available information embedded in data. The mathematical complexity of the system needs to be simplified, and one way is a lumped parameter model where the partial differential equation is replaced by a finite set of ordinary differential equations. In the classical situation when the model structure is based on physical knowledge and any information from available data is not taken into account, this approach is called white box modelling. The opposite of white box models are black box models that are exclusively based on data, where prior physical information is not taken into account. Benefiting from both modelling approaches, the grey box approach combines the information from the prior physical knowledge and information embedded in the data. In the grey box modelling framework it is possible to give direct physical interpretation of the estimated parameters. This paper introduces the grey box modelling approach and suggests a maximum likelihood method for parameter estimation where the likelihood function is evaluated using a Kalman filter technique. The model and model parameters are validated by applying statistical methods using all the available data.

1 Introduction

Traditionally, in groundwater well field hydrological modelling the spatio- temporal variation is described by deterministic partial differential equations with several input and output variables, e.g. discharge from the wells and the piezometric heads at the boundaries. However, for many practical applications, like

[1]DTU Informatics, Technical Unversity of Denmark; Richard Petersens Plads (bg. 305), DK-2800 Kgs. Lyngby, Denmark
[2]DHI, Agérn Allé 5, DK-2970 Hørsholm, Denmark

those connected to control, optimization and forecasting, it is essential to reduce the complexity of the mathematical expressions and to enable a rigorous stochastic description of the dynamics.

A popular approach for simplification is to consider a lumped parameter model where the partial differential equation is replaced by a finite set of ordinary differential equations. It is convenient to use a state-space model formulation of the ordinary differential equations, which then introduces a set of state space variables describing the dynamics of the considered system. In the classical approach the state-space model is formulated using all the available physical information, i.e. the known physical characteristics and well-established models of subprocesses. This modelling approach is often termed white box modelling, since all aspects of the model are formulated using prior physical knowledge and since any information embedded in observations is disregarded. A serious drawback of the classical approach is the difficulties involved with obtaining a reasonable parametrization. Generally the total model has a rather large number of parameters, and, due to the unavoidable idealizations, simplifications and unknown parameters, introduced both into the models of each of the individual subprocesses and into the coupling between the various subprocesses, it is very difficult to predict the accuracy of the total model.

For the opposite approach, which often is termed a black box approach, the model is based on groundwater well field data and statistical methods. This implies that both the model structure and the parametrization is deduced and validated by applying a series of statistical methods using all the available data. The use of statistical methods also enables a possibility for using rigorous stochastic dynamical models, which then provide methods for predicting the uncertainty of the model predictions. Since the well field data are sampled at discrete times, the model is most frequently formulated in discrete time as a difference equation. However, one serious drawback of the discrete time formulation is that information about the physical parameters is partially hidden in the discrete time parametrization. It is most often impossible, based on a discrete time formulation, to find a reasonable continuous time model, due to observational errors, embedding problems, or limitations in the flexibility of the model. If it is impossible to obtain a suitable continuous time formulation, it is also impossible to change the sampling time properly. Hence, it is desirable to use a formulation and an estimation method, where the parametrization is kept in continuous time. Furthermore, a continuous time stochastic model ensures a more reasonable physical interpretation of the parameters, and it allows to use the knowledge of e.g. physical constants or balance relations to improve the parametrization. Finally, if the estimation takes place in continuous time, information about the uncertainty due to quantization of physical characteristics may appear directly as a part of the estimation procedure.

The suggested grey box approach consists of models which are stochastic state-space models consisting of a set of stochastic differential equations (SDE's) (*Øksendal*, 2003) describing the dynamics of the system in continuous time and a set of observation equations in discrete time. Grey box modelling enables efficient model building, which gives a powerful method for combining prior physical knowledge regarding the system with information embedded in data series. In this paper, this approach will be introduced and applied for groundwater models, with emphasis on modelling of groundwater well fields. The main modelling framework will be illustrated, as well as the suggested maximum likelihood estimation method where the likelihood function is evaluated using an extended Kalman filter. Finally, a simple example will be given to illustrate a simple grey box approach for a single well that can be extended to consider other neighboring wells, and, in the near future, hopefully describe the spatio-temporal variation of the well field.

2 Grey box approach for groundwater modeling

Translating the ordinary differential equation (ODE) model into a stochastic state-space model is often a rather straightforward procedure, where the SDE models are replaced with the ODE models with the addition of one or more algebraic equations describing how measurements are obtained at discrete time instants. Most often the models are formulated as continuous-discrete time state-space models and in its most general form it is written as

$$d\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta})dt + \boldsymbol{\sigma}(t, \boldsymbol{\theta})d\boldsymbol{\omega} \tag{1}$$

$$\boldsymbol{y}_k = \boldsymbol{h}(\boldsymbol{x}_k, \boldsymbol{u}_k, t_k, \boldsymbol{\theta}) + \boldsymbol{e}_k \tag{2}$$

where $t \in \mathbb{R}$ is time ($t_k, k = 0, \ldots, N$ are sampling instants); $\boldsymbol{x}_t \in \mathbb{R}^n$ is a vector of state variables; $\boldsymbol{u}_t \in \mathbb{R}^m$ is a vector of input variables; $\boldsymbol{y}_k \in \mathbb{R}^l$ is a vector of output variables; $\boldsymbol{\theta}_t \in \mathbb{R}^p$ is a vector of possibly unknown parameters; $\boldsymbol{f}(\cdot) \in \mathbb{R}^n$, $\boldsymbol{\sigma}(\cdot) \in \mathbb{R}^{n \times n}$ and $\boldsymbol{h}(\cdot) \in \mathbb{R}^l$ are nonlinear functions; $\{\boldsymbol{\omega}\}$ is a n-dimensional standard Wiener process, and $\{\boldsymbol{e}_k\}$ is a l-dimensional white noise process with $\boldsymbol{e}_k \in N(\boldsymbol{0}, \boldsymbol{S}(\boldsymbol{u}_k, t_k, \boldsymbol{\theta}))$. The standard Wiener process is a continuous stochastic process with stationary and independent Gaussian time increments, which have the mean value zero and a covariance $\boldsymbol{S}$ equal to the magnitude of the increments (*Jazwinski*, 1970).

The first term on the right side of equation (1) is called the drift term, and the second term the diffusion term. The drift is the descriptive term, representing the physical structure of the system, Any prior physical knowledge regarding the model can be included, since the parameters in the ODE models usually provide some physical interpretation of the system. Furthermore, most experts

in hydrogeology are familiar with this way of modelling groundwater flow, where all parameters provide the model some physical assessment.

The diffusion term of the SDE model is considered to provide a suitable interpretation of the errors that exist due to the fact that the mathematical model is often not describing the true process exactly. However, the gap between the true process and the model should be reduced and by estimating the diffusion in the model, any unrecognized phenomena or un-modelled inputs can be found and considered in the model (*Madsen and Holst*, 2000). The most serious lack is related to some specific state description in the model, and by extending this particular state description by additional state variables, more generic methods for systematic improvement of the model structure is attained.

The observation equation (2) relates the discrete time observations to the state variables at time points where observations are available. When determining unknown parameters of the model from a set of data, the model equations in (1) and (2) enables the model with flexibility consisting of possibilities for varying sample times and missing observations in the data series. The model provides a separation between the process noise and the measurement noise, which allow the parameters to be estimated in a prediction error setting using statistical methods, like the maximum likelihood, which is introduced in the next section.

3 Parameter estimation and model validation

In hydrological modelling two estimation methods are often used, the Output Error method (OE) and the Prediction Error method (PE). The OE method minimizes the sum of squared deviation between model simulation and corresponding observations, whilst the PE method minimizes the sum of squared one step prediction error. Comparison shows that for simulations the two methods perform quite similar, but the estimated parameters are less biased with the PE method. Furthermore, uncertainty information is provided by the PE method, for which gives an advantage in short-term predictions (*Jonsdottir et al.*, 2006). The maximum likelihood method presented below is a PE method.

Given the model structure in (1) and (2), the unknown parameters can be determined by finding the parameters that maximize the likelihood function of a given sequence of measurements, i.e. by Maximum Likelihood method. The rule $P(A \cap B) = P(A|B)P(B)$ is applied to express the likelihood function as a product of conditional densities, and by representing the measured sequence

by $\mathcal{Y} = [\boldsymbol{y}_k, \ldots, \boldsymbol{y}_0]$ the likelihood function is the joint probability density

$$L(\boldsymbol{\theta}; \mathcal{Y}_N) = P(\mathcal{Y}_N|\boldsymbol{\theta}) = \left(\prod_{k=1}^{N} P(\boldsymbol{y}_k|\mathcal{Y}_{k-1}, \boldsymbol{\theta}) \right) P(\boldsymbol{y}_0|\boldsymbol{\theta})$$

To obtain an exact estimation of the likelihood function, the continuous-discrete filtering problem needs to be solved, and the initial probability density function $P(\boldsymbol{y}_0|\boldsymbol{\theta})$ must be known and parameterized, and all subsequent conditional densities must be determined to successively solve Kolmogorov's forward equation (*Kloeden and Platen*, 1999). In practice, however, this approach is not computationally feasible and an alternative is required. Since the SDE's in (1) are driven by a Wiener process, which has Gaussian increments, the conditional densities can be approximated by Gaussian densities. For linear models the Kalman filter provides the exact solution for the filtering problem, and for nonlinear models the problem is approximated by applying the extended Kalman filter (*Madsen et al.*, 2004).

The Gaussian density is completely characterized by its mean and covariance, which are denoted by $\hat{\boldsymbol{y}}_{k|k-1} = E\{\boldsymbol{y}_k|\mathcal{Y}_{k-1}, \boldsymbol{\theta}\}$ and $\boldsymbol{R}_{k|k-1} = V\{\boldsymbol{y}_k|\mathcal{Y}_{k-1}, \boldsymbol{\theta}\}$, respectively, and by introducing an expression for the innovation $\boldsymbol{\epsilon}_k = \boldsymbol{y}_k - \hat{\boldsymbol{y}}_{k|k-1}$ the likelihood function can be rewritten as

$$L(\boldsymbol{\theta}; \mathcal{Y}_N) = \left(\prod_{k=1}^{N} \frac{\exp\left(-\frac{1}{2}\boldsymbol{\epsilon}_k^\top \boldsymbol{R}_{k|k-1}^{-1} \boldsymbol{\epsilon}_k\right)}{\sqrt{\det(\boldsymbol{R}_{k|k-1})}\left(\sqrt{2\pi}\right)^l} \right) P(\boldsymbol{y}_0|\boldsymbol{\theta})$$

where the conditional mean and covariance are calculated using an Extended Kalman Filter. Finally, the parameter estimates can be determined by conditioning on the initial values and solving the optimization problem

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \Theta} \{-\ln(L(\boldsymbol{\theta}; \mathcal{Y}_N|\boldsymbol{y}_0))\}$$

With the unknown parameters of the model estimated by the ML method, along with corresponding standard deviations, statistical tests can be performed to check if the parameters are significantly different from zero, which may indicate that some improvement is needed for the model structure.

One of the important aspects of the modelling framework is its predictive ability, which implies that the prediction errors are examined for any systematic pattern for further extension of the model. Most importantly, the sample autocorrelation function and sample partial autocorrelation function of the residuals are investigated to detect if two or more consecutive residuals are dependent or, in contrast, can be regarded as white noise (*Kristensen et al.*, 2004, *Madsen*, 2008). Correlation between the residuals indicates that the model is not adequate.

For the groundwater well field there is strong relation between any two or more neighboring wells, i.e. withdrawal from one well is monitored also in the other wells. Relating the model more on the measured piezometric heads in the wells and apply statistical methods to improve the model structure, a model is obtained which is adjusted to the objectives and more able to cope with real-time measurements. A model for a single well in the field can, therefore, easily be extended to consider other wells in the surroundings and, eventually, the entire well field.

4 Grey box well field modeling: a simple test case

By discharging water from one particular well does most often have some influence on the other wells in the surroundings. By pumping from well no.1 clearly affects the water elevation in well no.2. A lumped parameter model is formulated to see if well no.2 can be totally described by well no.1. The grey box model for the relation between the wells is

$$dh_1 = \left[\frac{T_{1-\infty}}{C_1} (h_\infty - h_1) + \frac{T_{1-12}}{C_1} (h_{12} - h_1) - \frac{Q}{C_1} \right] dt + \sigma_1 d\omega_1 \tag{3}$$

$$dh_{12} = \left[\frac{T_{1-12}}{C_{12}} (h_1 - h_{12}) + \frac{T_{12-\infty}}{C_{12}} (h_\infty - h_{12}) + \frac{T_{12-2}}{C_{12}} (h_2 - h_{12}) \right] dt + \sigma_{12} d\omega_{12} \tag{4}$$

$$dh_2 = \left[\frac{T_{12-2}}{C_2} (h_{12} - h_2) + \frac{T_{2-\infty}}{C_2} (h_\infty - h_2) \right] dt + \sigma_2 d\omega_2 \tag{5}$$

where h_1 and h_2 are state variables for the pressure heads in the operating well and the well of interest, respectively, and h_{12} is a state variable for the water table elevation between the two wells, whilst h_∞ is the static water table elevation. The parameters T and C in the model represent the hydraulic transmissivity from one state variable to the next and lumped parameters for the storage coefficient of the aquifer, respectively. Q is the discharge rate from well no.1, kept at 90 m^3/min until it is turned off.

The pressure head in well no.2 is observed, so the observation equation for well no.2 is

$$H_{2,k} = h_{2,k} + e_k \tag{6}$$

If the elevation in the observed well was exclusively depending on the discharge from well no.1, the residual analysis would clarify that no autocorrelation exist. However, Figure 1 shows that the residuals cannot be considered white noise, and indicate that some important features are not counted for in the model structure. Therefore, the observed well cannot be exclusively de-

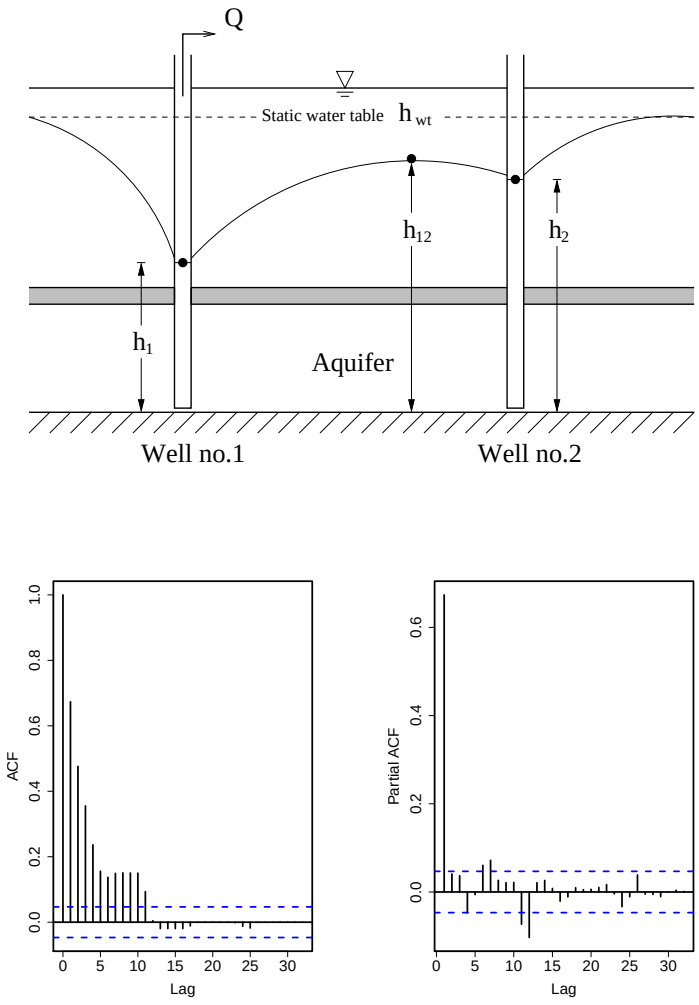

Figure 1: The conceptual model, autocorrelation and partial autocorrelation for the residuals from well no.2.

scribed by the discharge in well no.1, and the model structure needs to be improved by taking into account a new state variable, e.g. leakage coefficient, other neighboring wells or possible boundaries in the well field. Thus, from this simple case the model can be extended systematically, such that the essential prior physical knowledge regarding the whole groundwater well field is combined with available information to get more accurate results to describe the spatio-temporal variation of the well field.

5 Conclusion

A grey box approach of groundwater flow models has been presented, which combines the best from deterministic and stochastic modelling for identification of models for model-based control of groundwater well fields. The grey box approach is based on flexible and statistical methods for continuous-discrete time stochastic state-space models, which are just as appealing as ordinary differential equation models from an engineering point of view. One of the most important aspects of the approach is its constructive features for performing model validation by means of statistical tests and residual analysis, where the significance of parameters may provide information about the validity of the proposed model. Based on these methods it has been demonstrated that the rather simple grey box model can be extended towards an operational description of the spatio-temporal variation of the groundwater well field.

References

Jazwinski AH (1970) *Stochastic processes and filtering theory*. Academic Press, New York, USA.

Jonsdottir H, Madsen H, Palsson OP (2006) Parameter estimation in stochastic rainfall-runoff models. *Journal of Hydrology* **326**:379-393.

Kloeden P, Platen E (1999) *Numerical Solutions of Stochastic Differential Equations*. Springer-Verlag.

Kirstensen NR, Madsen H, Jørgensen SB (2004) A method for systematic improvement of stochastic grey-box models. *Computers and Chemical Engineering* **28**:1431-1449.

Madsen H, 2008 *Time series analysis*. Chapman & Hall/CRC.

Madsen H, Holst J (2000) *Modelling non-linear and non-stationary time series*. Technical University of Denmark, Informatics and Mathematical Modelling.

Madsen H, Nielsen JN, Lindström E, Baadsgaard M, Holst J (2004) *Statistics in finance*. Lund University, Centre for mathematical sciences.

Øksendal B (2003) *Stochastic differential equations; an introduction with applications* (6th ed.). Springer.

PAPER B

Predictions for groundwater well fields using stochastic modelling

Authors:
F. Ö. Thordarson, H. Madsen, H. Madsen

In preceedings:
HydroPredict (2010)

Predictions for Groundwater Well Fields using Stochastic Modeling

Fannar Örn Thordarson[1], Henrik Madsen[1], Henrik Madsen[2]

Abstract

A stochastic modelling framework for identifying groundwater well fields is presented, which combines prior physical knowledge of dynamic groundwater well field systems with available information embedded in data. The model is a conceptual stochastic model, formulated in continuous-discrete time state space form that facilitates a direct physical interpretation of the estimated parameters. The parameter estimation method is a maximum likelihood method, and the model parameters are validated by applying statistical methods using all the available data. The statistical tools are used to identify the deficiencies in a model that is considered too simple. Even though the predictions seem adequate, statistical methods show that the model needs to be extended to be able to provide accurate predictions for the groundwater level in all wells.

Key words:
Groundwater, Well field model, Stochastic differential equations, Grey box model, Prediction, Parameter estimation, Maximum likelihood method

1 Introduction

It is essential to ensure high quality drinking water in the future, which then calls for reliable operation and management of the groundwater resources at well fields. One of the foundations of the groundwater resource management is the mathematical model that describes the behavior of the aquifer penetrated by one or several wells. For control, optimization and forecasting, the complexity of the mathematical expressions needs to be reduced to enable more rigid stochastic representation of the dynamics.

[1]DTU Informatics, Technical Unversity of Denmark; Richard Petersens Plads (bg. 305), DK-2800 Kgs. Lyngby, Denmark
[2]DHI, Agérn Allé 5, DK-2970 Hørsholm, Denmark

The groundwater elevation in the well field varies in both time and space and is traditionally described by the governing equation for groundwater flow, which most frequently is facilitated by a deterministic partial differential equation (*Anderson and Woessner*, 2002). With multiple discharge locations in the well field the utility of the governing equation becomes highly complex. A popular approach for simplification is to consider a lumped parameter model where the partial differential equation is replaced by a finite set of ordinary differential equations in state-space form, which then introduces a set of state-space variables describing the dynamics of the well field. The state-space model is formulated using all the available prior physical knowledge, which include the known physical characteristics of the considered system and any auxiliary processes connected to the well field. This approach disregards any stochasticity related to the variation in time and space with a serious drawback of obtaining a reasonable parametrization. The total model is often characterized by having a large number of parameters and due to unavoidable idealizations, simplifications and unknown parameters, it is difficult to predict the accuracy of the total model. This modelling approach is often referred to as a white-box approach, since the model structure is completely transparent and the variation in the available data is neglected.

On the contrary is the black-box approach where the model is formulated by only considering the available well field data and statistical methods are applied to reduce and validate the structure and the parametrization for the groundwater well field. The used of statistical methods enables a possibility for using rigorous stochastic dynamical models which then provide methods for predicting the uncertainty of the model predictions. However, the data is sampled at discrete time and a drawback of the discrete time formulation is that information about the physical parameters is partially hidden, and due to measurement errors or limitations in model flexibility, a reasonable continuous time model cannot be obtained.

It is desirable to obtain a modelling approach that reduces the gap between the conventional models based on physical characteristics and the pure statistical discrete time approach. Using formulation and estimation method, where the parametrization is kept in continuous time, a continuous time stochastic model is obtained where the estimated parameters do have a direct physical interpretation. Hence, in relation to the well field model any knowledge of physical constants and water balance relations can be exploited to improve the parametrization. This modelling approach is referred to as the grey-box approach, since being a combination of the other two approaches.

This paper presents a formulation and estimation of a simple continuous time stochastic model for the groundwater well field that explicitly describes how the measurements and model errors enter into the model, and, due to contin-

uous time formulation, the model facilitates a direct physical interpretation of the estimated parameters. Based on the proposed method it is demonstrated that the rather simple continuous time stochastic model constitutes an operational description of the spatio-temporal variation for simulations and predictions for the considered groundwater well field.

2 Continuous-Time Stochastic Model for Groundwater Well Field

By considering the lumped parameter approach in state-space form, represented by a finite set of ordinary differential equations (ODEs), the translation into a set of stochastic differential equations (SDEs) is often a rather straightforward procedure. This is usually obtained by replacing the ODE models with the SDE models, which in addition also includes one or more algebraic equations describing how measurements are obtained at discrete time instants. Most often the models are formulated as continuous-discrete time state-space models and in its most general form it is written as

$$d\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, t; \boldsymbol{\theta})\, dt + \boldsymbol{\sigma}(t, \boldsymbol{u}_t; \boldsymbol{\theta})\, d\boldsymbol{\omega}_t \tag{1}$$

$$\boldsymbol{y}_k = \boldsymbol{h}(\boldsymbol{x}_k, \boldsymbol{u}_k, t_k; \boldsymbol{\theta}) + \boldsymbol{e}_k \tag{2}$$

where $t \in \mathbb{R}_0$ is time ($t_k, k = 1, \ldots, N$ are sampling instants); $\boldsymbol{x}_t \in \mathbb{R}^n$ is a vector of state variables; $\boldsymbol{u}_t \in \mathbb{R}^m$ is a vector of input variables; $\boldsymbol{y}_t \in \mathbb{R}^l$ is a vector of output variables; $\boldsymbol{\theta} \in \mathbb{R}^p$ is a vector of unknown parameters; $\boldsymbol{f}(\cdot) \in \mathbb{R}^n$, $\boldsymbol{\sigma}(\cdot) \in \mathbb{R}^{n \times n}$ and $\boldsymbol{h}(\cdot) \in \mathbb{R}^l$ are nonlinear functions; $\{\omega\}$ is a n-dimensional standard Wiener process, and e_k is a l-dimensional white noise process with $e_k \sim N(0, \boldsymbol{S}(\boldsymbol{u}_t, t_k, \boldsymbol{\theta}))$. The standard Wiener process is a continuous stochastic process with stationary and independent Gaussian time increments, which have the mean value zero and a covariance $\boldsymbol{S}$ equal to the magnitude of the increments (*Jazwinski*, 1970). Equation (1) is called the system equation, whereas equation (2) is the observational equation.

The first term on the right side of the system equation is usually called the drift term, since it represents the physical structure of the system, determined and formed from the system of ordinary differential equations. Hence, any prior physical knowledge regarding the physical structure is included in the drift term where the parameters provide some physical interpretation of the system. Furthermore, the physical characteristics of the drift term are expressions most hydrogeologists are familiar with from formulating the traditional groundwater flow models.

The second term on the right side of the system equation is the diffusion term

of the SDE model, which provides a suitable interpretation of the errors that exist due to the fact that the mathematical model is often not describing the true process exactly. However, the gap between the true process and the model should be reduced and by estimating the diffusion in the model, any unrecognized phenomena or unmodelled inputs can be detected and directly or indirectly considered in the model. Frequently is this discrepancy related to some specific state description in the model, and by extending this particular state description by additional state variables, more generic methods is obtained for systematic improvement of the model.

The observation equation (2) then relates the discrete time observations to the state variables at time points where observations are available. When determining unknown parameters of the model from a set of data, the model equations in (1) and (2) enables flexible estimation that can account for varying sample times and missing observations in the data series. The model provides a separation between the process noise and the measurement noise, which allow the parameters to be estimated in a prediction error setting, using statistical methods and the maximum likelihood method.

3 Parameter Estimation

A solution to the well field prediction problem is to optimize a set of parameters, such that the model for the groundwater levels in the well field sufficiently fits the available data. The most direct terminology is to minimize the error between the model output and the observed output for the well field. For such an objective, mainly two estimation methods have been applied for optimizing the parameters in hydrological studies; the Output Error method (OE) and the Prediction Error method (PE).

The OE method minimizes the sum of squared simulation error and is applied for white-box models with well described physical characteristics, without considering variation in the available data. The parameters estimated by the OE method are, in general, not provided with any uncertainty. Furthermore, the OE method can only be considered for offline estimation, i.e. the estimates are only depending on the initial values; for online estimation the state estimates are updated for every time instants. The PE method seeks for minimizing the sum of squared one-step prediction error to obtain the best fitted model for the groundwater level in the well field, and the PE method includes both offline and online estimation. Moreover, the PE method also provides an uncertainty for the estimates, which is well suited for short-term predictions.

Given the model structure in (1) and (2), the unknown parameters can be de-

termined by finding the parameters that maximize the likelihood function of a given sequence of measurements, i.e. by the Maximum Likelihood (ML) method. From probability theory the rule of independent probabilities can be applied to express the likelihood function as a product of conditional densities, and by representing the measured sequence by $\mathcal{Y}_K = [\boldsymbol{y}_K, \cdots, \boldsymbol{y}_0]$ the likelihood function is the joint probability density

$$L(\boldsymbol{\theta}; \mathcal{Y}_K) = p(\mathcal{Y}_K|\boldsymbol{\theta}) = \left(\prod_{k=1}^{K} p(\boldsymbol{y}_k|\mathcal{Y}_{k-1}, \boldsymbol{\theta}) \right) p(\boldsymbol{y}_0|\boldsymbol{\theta}).$$

To obtain an exact estimation of the likelihood function, the continuous-discrete filtering problem needs to be solved, and the initial probability density function $p(\boldsymbol{y}_0|\boldsymbol{\theta})$ must be known and parameterized, and all subsequent conditional densities must be determined to successively solve Kolmogorov's forward equation (*Kloeden and Platen*, 1999). In practice, however, this approach is not computationally feasible and an alternative is required. Since the SDE's in (1) are driven by a Wiener process, which has Gaussian increments, the conditional densities can be approximated by Gaussian densities. For linear models the Kalman filter provides the exact solution for the filtering problem, and for nonlinear models the problem is approximated by applying the extended Kalman filter (*Madsen et al.*, 2004).

The Gaussian density is completely characterized by its mean and covariance, which are denoted by $\hat{\boldsymbol{y}}_{k|k-1} = \mathrm{E}\{\boldsymbol{y}_k|\mathcal{Y}_{k-1}, \boldsymbol{\theta}\}$ and $\boldsymbol{R}_{k|k-1} = \mathrm{V}\{\boldsymbol{y}_k|\mathcal{Y}_{k-1}, \boldsymbol{\theta}\}$, respectively, and by introducing an expression for the innovation $\boldsymbol{\epsilon} = \hat{\boldsymbol{y}}_{k|k-1} - \boldsymbol{y}_k$ the likelihood function can be rewritten as

$$L(\boldsymbol{\theta}; \mathcal{Y}_N) = \left(\prod_{k=1}^{N} \frac{\exp\left(-\frac{1}{2}\boldsymbol{\epsilon}_k^\top \boldsymbol{R}_{k|k-1}\boldsymbol{\epsilon}_k\right)}{\sqrt{(}\det(\boldsymbol{R}_{k|k-1}))\sqrt{(2\pi)^l}} \right) p(\boldsymbol{y}_0|\boldsymbol{\theta})$$

and thereof, the parameter estimates can be determined by conditioning on the initial values and solving the optimization problem

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \left(\log(\boldsymbol{\theta}; \mathcal{Y}_N|\boldsymbol{y}_0)\right).$$

With the unknown parameters of the model estimated by the ML method, along with corresponding standard deviations, statistical tests can be performed to check if the parameters are significantly different from zero, which then indicates that some improvement is needed for the model structure. The parameters of the diffusion term in equation (1) are included in the ML estimation.

One of the main aspects of the modelling framework is its predictive ability, which implies that the output errors are examined for any systematic pattern

for further extension of the model, as well as investigation of the sample autocorrelation function and sample partial autocorrelation function of the residuals to detect if two or more consecutive residuals are dependent or, in contrast, can be regarded as white noise (*Kristensen et al.*, 2004). Correlation between the residuals indicates that the model is not adequate for prediction, since systematic errors are detected in the model that can affect the model prediction drastically. An adequately parameterized model is characterized by having uncorrelated residuals (*Madsen*, 2008).

4 An Example

The following is an example to illustrate the important features of the continuous time stochastic model described above; the lumped model for the well field, the parameter estimation and model prediction. The well field has three pumping wells, which all pump from the same aquifer. These three wells are a part of water distribution network with 21 operating wells attached, where all wells are pumping from the same aquifer. The total well field is divided into three groups due to geographical location. Here, one of these is studied.

The conceptual model is sketched in Figure 1a, showing the three wells located on a straight line, that is, well No. 2 is located on the line between well No. 1 and well No. 3. This simplifies the model by assuming that drawdown in well No. 3 when pumping from No. 1 is detected in well No. 2 as well. This assumption is also valid when the water level changes in well No. 1 when pumping from well No. 3.

The objective is to predict the piezometric heads in the wells when pumping from a confined aquifer. However, since the lumped parameter model is considered for the model structure, the parameters are lumped vertically, from datum to the piezometric head, and the suggested model for the groundwater well field is expected to consist of a number of reservoirs where the water levels in the reservoirs are the state variables in the state-space representation (*Jacobsen et al.*, 1997). As illustrated in Figure 1b, the only measured state variables are the water-levels in the wells. The water levels between any two wells, and at the boundaries, are unobserved state variables, which will be estimated in relation to the observations in its two adjacent operating wells. The behavior of the water table between two operating wells is nonlinear, but by assuming several linear reservoirs for the water table to represent the flow from one well to another, the water table can be linearly approximated. The water level, or the reservoirs, in the unobserved states does never dry out, indicating that at least one of the unobserved reservoirs between every two observed wells is infiltrated with additional water.

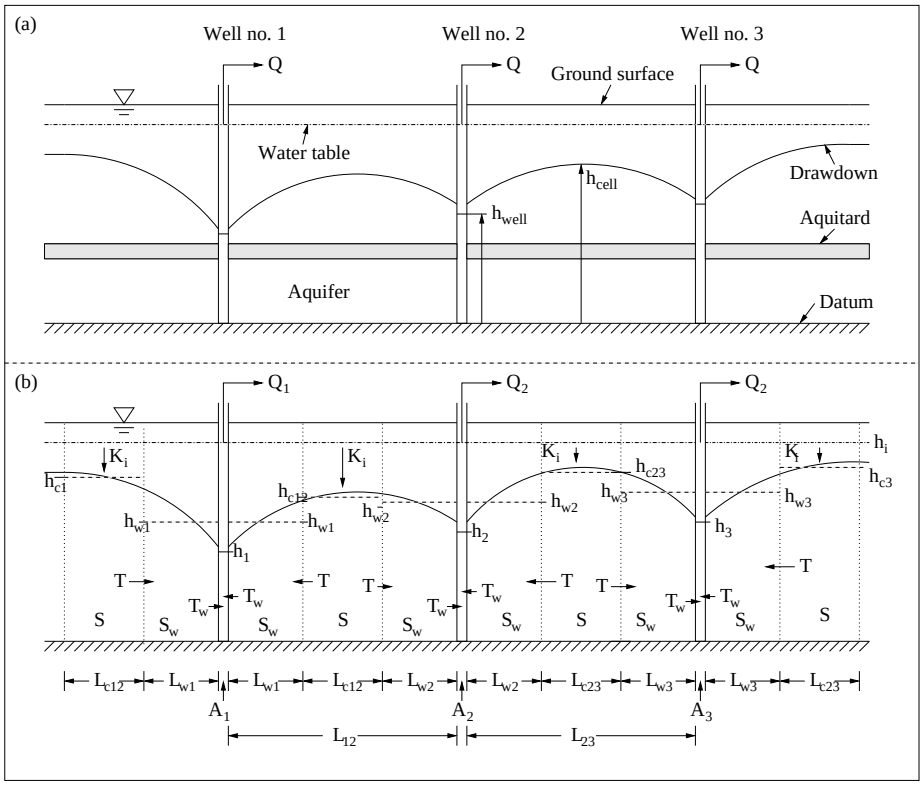

Figure 1: Conceptual model for a well field with 3 operating wells: (a) The classical illustration of the model. (b) The lumped model represented as number of linear reservoirs.

Considering the states as given in Figure 1b, and with the index i indicating the state of interest, the suggested stochastic state space model, as in equation (1), is represented as follows: The pumping wells are the observed states (h_3, h_7 and h_{11} in Figure 1b) and their dynamics are described as

$$dh_{i,t} = \left[\frac{K}{A_i}h_{i-1,t} - \frac{2K}{A_i}h_{i,t} + \frac{K}{A_i}h_{i+1,t} - \frac{1}{A_i}Q_{i,t}\right]dt + \sigma_i d\omega_{i,t}$$

with K assumed to be the lumped hydraulic conductivity and A_i is considered as the areal closest to the well directly affected by the discharge. Here, and in all the following system equations for the well field, the σ_i values are constants representing the variation of the system noise for state description i, where $i = 1,\ldots,13$, and corresponding noise term $d\boldsymbol{\omega}_i$ is assumed to be independent standard Wiener process, and also assumed independent from the measurement noise in the observation equation.

The state variables illustrating the recharge of the aquifer between operating wells (h_5 and h_9) become

$$dh_i = \left[\frac{K}{SL_i}h_{i-1,t} - \frac{1}{SL_i}(2K + \frac{1}{K_{inf}})h_{i,t} + \frac{K}{SL_i}h_{i+1,t} - \frac{h_{inf}}{SL_iK_{inf}}Q_{i,t}\right]dt + \sigma_i d\omega_{i,t}$$

The same goes for the recharged boundary states (h_1 and h_{13}), except for one term is neglected in each case; for h_1 the first term in the square brackets is omitted, and for h_{13} the last term inside the square brackets. S is the storage coefficient for the lumped flow and L_i is the estimated size of reservoir i. h_{inf} is the estimated boundary condition, i.e. the water level approaches the undisturbed water table if no pump is active in the well field for a reasonably long time. The term K_{inf} is the estimated resistance for the flow from the boundaries to the reservoirs.

For all the remaining states, the intermediate states of the water level in the reservoirs is represented as

$$dh_i = \left[\frac{K}{SL_i}h_{i-1,t} - \frac{2K}{SL_i}h_{i,t} + \frac{K}{SL_i}h_{i+1,t}\right]dt + \sigma_i d\omega_{i,t}.$$

There are three observation equations since there are three measured water levels in the wells, i.e.

$$\begin{aligned} Y_{1,k} &= h_{3,k} + e_{1,k} \\ Y_{2,k} &= h_{7,k} + e_{2,k} \\ Y_{3,k} &= h_{11,k} + e_{3,k} \end{aligned}$$

where the e_1, e_2, e_3, correspond to the measurement noises.

The parameter estimation is shown in Table 1. The estimation for the hydraulic conductivity and the storage coefficient are reasonably estimated, but compared to results from a pumping test for the aquifer the estimates are orders of magnitudes higher. This is explained by the fact that these two estimated parameters are lumped vertically and correspond to routing of water and storage in the aquifer, as well as all the layers above it. Therefore, it is impossible to compare results from pumping tests and the lumped estimates. The two estimated values, K and S, correspond to the individual reservoir in the lumped model, where K is assumed as routing coefficient per length unit and the storage S is considered as the total storage per length unit in each reservoir. Model extension that takes consideration to the different layers in the conceptual model can be implemented into the introduced stochastic model, but no such attempt is made in this particular study.

By performing t-tests the parameters can be checked for being significantly different from zero. Status for significance of each parameter is displayed in the

last column in Table 1, and it shows that the variances for the system noises, regarding the boundary conditions, are not significant (σ_1 and σ_{13}). For extending the model further, focus should be on the state descriptions for the boundary conditions, since from the parameter estimation it can be concluded that these states are not entirely fulfilled with the present description. The most probable cause is related to the other two groups of wells in the total well field and to get a better understanding of the boundary conditions for this small group of three wells, correlation to the other groups need to be exploited.

It is interesting to see how adequate the model is to predict the water level in the three wells. Figure 2 displays a comparison between the observations (solid line) and corresponding model output (dashed line) for the pumping wells. Although it appears as the prediction follows the observations rather well, there is a clear difference for all three wells where the greatest deviation is in relation to abrupt changes in the water level, i.e. when a pump is switched on or off. Figure 3 shows that the difference between the model and the observations is serially correlated, which indicates that an improved model should be obtained by addition of a reservoir between operating wells.

This example shows how the presented lumped stochastic model can be used for parameter estimation and prediction for a groundwater well field. It is

Table 1: Estimated values for several parameters in the stochastic well field model. K:[m/min], S [-]; A_i [m^2]; h_{inf} [m].

Parameter	P. test	θ	std(θ)	Significant
K	0.0420	1.09	0.03	YES
S	0.0012	2.08	0.36	YES
A_3	-	10.25	0.71	YES
A_7	-	5.48	0.29	YES
A_{11}	-	6.26	0.53	YES
h_{inf}	-	7.14	0.36	YES
σ_1	-	0.04	0.03	NO
σ_3	-	0.36	0.05	YES
σ_5	-	0.03	0.02	YES
σ_7	-	0.29	0.02	YES
σ_9	-	0.17	0.02	YES
σ_{11}	-	0.21	0.02	YES
σ_{13}	-	0.02	0.01	NO
S_1	-	0.00	0.00	NO
S_2	-	0.00	0.00	NO
S_3	-	0.00	0.00	NO

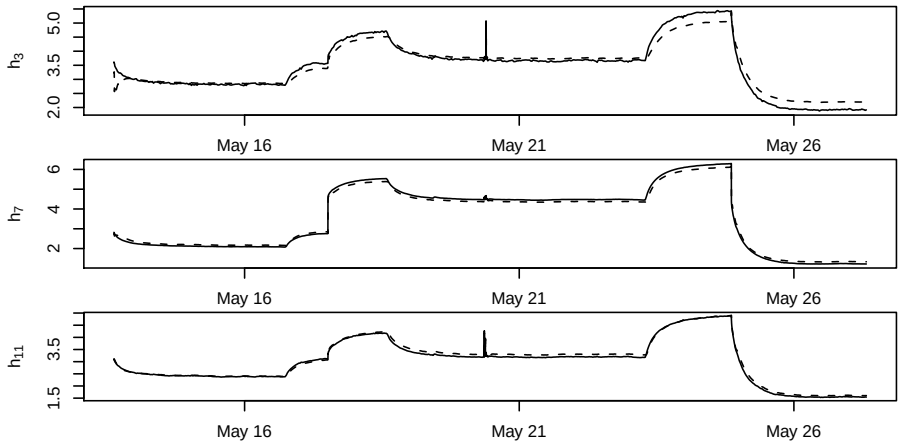

Figure 2: Comparison between measurements (solid line) and predictions (dashed line) for all operating wells.

also shown how statistical methods can be applied to detect deficiencies in a model, as well as suggest which state descriptions require improvement. By optimizing the parameters with the ML method, the model is able to predict the water levels in the wells within the limited region, but from a statistical point of view an improved model is need to obtain more adequate results.

5 Conclusion

A continuous time stochastic model for a groundwater well field has been presented. This modelling framework combines the best from deterministic and stochastic modelling for identification of models, for model-based control of groundwater well fields. The model basis are the state descriptions in the stochastic state-space model, derived from stochastic differential equation models, which are just as appealing as ordinary differential equation models from an engineering point of view. The maximum likelihood method provides uncertainty to the estimates, which is highly important for performing model validation by means of statistical tests and residual analysis. Based on these methods it has been demonstrated that the rather simple stochastic model can be constructed to give sufficient results for the physically interpretable parameters. However, statistical tests showed that the model requires an extension to compose an operational description of the spatio-temporal variation of the groundwater well field, which eventually will improve the groundwater level predictions in the well field.

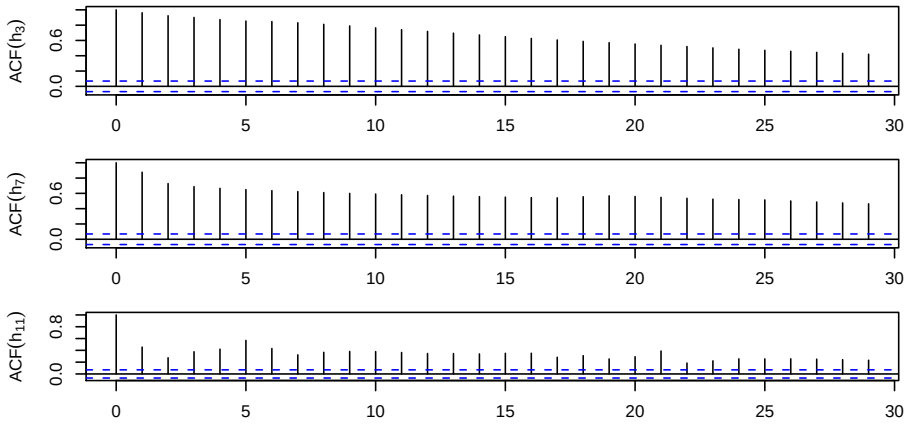

Figure 3: Autocorrelation functions for the residuals for all operating wells.

Acknowledgements

This work was partly funded by the Danish Strategic Research Council, Sustainable Energy and Environment Programme. For more information visit http://wellfield.dhigroup.com/.

References

Anderson MP, and Woessner WW (2002) *Applied groundwater modeling: simulation of flow and advective transport*. Academic Press, USA.

Jazwinski AH (1970) *Stochastic processes and filtering theory*. Academic Press, New York, USA.

Jacobsen JL, Madsen H, and Harremoës P (1997) A stochastic model for two-station hydraulics exhibiting transient impact. *Water Science and Technology* **36**(5):19-26.

Kristensen NR, Madsen H, and Jørgensen SB (2004) A method for systematic improvement of stochastic grey-box models. *Computers and Chemical Engineering* **28**:1431-1449.

Kloeden PE, and Platen E (1999) *Numerical solution of stochastic differential equations*. Springer.

Madsen H, Nielsen JN, Lindström E, Baadsgaard M, and Holst J (2004) *Statistics in finance*. Lund University, Centre for mathematical sciences.

Madsen H (2008) *Time series analysis*. Chapman & Hall/CRC.

PAPER C

Stochastic well field modelling using the grey box approach

Authors:
F. Ö. Thordarson, G. Dorini, H. Madsen, H. Madsen

Submitted to:
Advances in Water Resources (2012)

Stochastic well field modelling using the grey box approach

Fannar Örn Thordarson[1] Gianluca Dorini [1] Henrik Madsen[1] Henrik Madsen[2]

Abstract

The operation and management of a well field requires a reliable model for the groundwater flow to guarantee the robustness of the system. These models are usually physically-based, computationally intensive, and do not account for the various sources of uncertainty. Most often is the model structure highly complex, and needs to be reduced for the model to be feasible for control, optimisation and predictions of the system and the embedded uncertainty. By applying a grey box approach, a model is obtained that is operational for the well field management. The well field is presented by a system equation on a state space form where the states are represented by a set of stochastic differential equations, which consist of a drift term to describe the system dynamics in form of ordinary differential equations, and a diffusion term for the uncertainties in the model structure and the forcing of the system. Combining the system equation with a measurement equation for the observable states form the grey box model, which is simpler than the traditional physically-based groundwater flow models and identifiable from data. In this paper, the grey box model approach is described for groundwater flow in a confined aquifer that is penetrated by several pumping wells. A lumped parameter grey box model is introduced, and showing how this model can be improved; first, by extending the drift term in the model to account for missing and necessary dynamic behaviour in the physical system; and second, by extending the diffusion term to bound the prediction intervals of the model output. The maximum likelihood method is used for parameter estimation, and a quantile skill score criterion for performance evaluation, showing great improvements as the model is extended.

Key words:
grey box modelling, stochastic differential equations, groundwater flow, well field management, uncertainty assessment

[1]Informatics and Mathematical Modelling, Bldg. 305 DTU, DK-2800 Kgs. Lyngby, Denmark
[2]DHI, Agérn Allé 5, DK-2970 Hørsholm, Denmark

1 Introduction

This paper introduces grey box modelling of a well field that discharges groundwater from several wells, which all pump from the same confined aquifer. The grey box approach combines physical knowledge about the system with statistical modelling to obtain a stochastic model of the system. Physical knowledge provides a detailed structure of the system dynamics, often formulated as differential equations with physical parameters that can be estimated from literature, laboratory experiments or by model calibration. However, the physically-based model has approximation errors; which come from approximations regarding the model structure, uncertain model parametrisation, uncertainties in initial and boundary conditions that are not observable but have influence on the system, and deficient measurements (see discussion by, e.g., *Harremoës and Madsen*, 1999, *Rosbjerg and Madsen*, 2005, *Refsgaard et al.*, 2006).

Accuracy of the model is vital, especially for models that are utilised for predicting the future evolution of the physical system. Therefore, the use of statistical modelling prevents overparametrisation of the model (*Beven*, 1996). The physically-based models are often called white box models because of their transparency in the model structure. On the other hand, the statistical models do not have a structural identity, since the structure is exclusively based on the data and, hence, named black box models in contrast with the white box label of the physical models. The grey box approach bridges the gap between physical and statistical modelling, and facilitates a modelling framework in which prior physical information can be combined with information embedded in data (*Bohlin and Graebe*, 1995, *Harremoës and Madsen*, 1999, *Kristensen et al.*, 2004, *Bacher and Madsen*, 2011). Since the grey box model is determined from the physical structure of the system, it is expected that the model is somehow valid beyond the range covered by the measured data. Therefore, it is also expected that a grey box model is able to make better long-term predictions than black box models.

Groundwater modelling is widely applied in control, management and optimisation of aquifer systems (*Hansen et al.*, 2011, *Hendriks-Fransen et al.*, 2011, *Bauser et al.*, 2010, *Ahlfeld and Baro-Montes*, 2008, *Kollat and Reed*, 2006, *Siegfried and Kinzelbach*, 2006). To maintain feasibility in the control strategies for the pumps, the underlying model needs to be reliable. The key element of the introduced grey box model is the drift term of the model structure to describe the flow dynamics in the aquifer. The stepwise procedure from the governing equation for groundwater flow to formulation of the grey box model is illustrated in Section 2. This includes also the parameter estimation of the grey box model. In Section 3 the data, used in the following case study is described. The proposed models are discussed in detail in Section 4 and, also how the models

are related. Included are the estimation results for the model parameters and, eventually, the performance of the predictions of the models is evaluated and compared in order to decide on the most adequate model structure. The paper is then summarised with some concluding remarks in Section 5.

2 Stochastic well field model

The groundwater dynamics in a well field is described by the governing equation for groundwater flow (*Bear*, 1979):

$$S_S \frac{\partial h}{\partial t} = \nabla \cdot \boldsymbol{\kappa} \nabla h + R \tag{1}$$

where h [L] is the hydraulic head, $\boldsymbol{\kappa}$ [LT^{-1}] is the tensor matrix of the hydraulic conductivity, S_S [L^{-1}] is the specific storage and R [T^{-1}] represents any external stress affecting the groundwater flow. With given initial conditions and boundary conditions, and a given sequence for the stresses affecting the aquifer (R), the water level h can be determined by solving the governing equation in both time and space.

2.1 State space formulation

The governing equation (1) is a Partial Differential Equation (PDE), which can be solved by numerical methods; finite difference method, finite element method or finite volume method (see e.g. *Anderson and Woessner*, 2002, *Carrera*, 2008). To solve the groundwater flow equation numerically, the well field is discretised into a number of grid cells, i.e. the PDE is replaced by a finite set of Ordinary Differential Equations (ODEs). Hence, for the ODEs it is advantageous to use a state space formulation to describe the dynamics in the aquifer by a set of state space variables. Furthermore, a stochastic state space model includes a specific description of the measurement errors by the so-called measurement equation (*Madsen*, 2008).

To formulate the groundwater flow in a state space form, the governing equation (1) is initially approximated by discrete cells. By integration for each cell with respect to its volume, a representation of the water balance in cell i is obtained by including flow between cell i and all J neighbouring cells (the relation between the cells is sketched in Figure 1). Applying the divergence theorem (*Adams*, 1999) an ODE for cell i is attained (*Narasimhan and Witherspoon*, 1976,

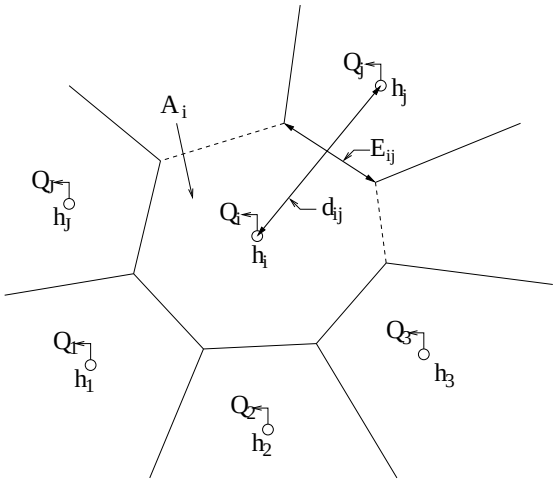

Figure 1: A sketch for cell i and the J neighbouring cells, with explanation for the parameters related to cell i.

Rozos and Koutsoyiannis, 2010), and is written

$$S_{S,i}V_i\frac{dh_i}{dt} = \sum_{j=1}^{J}\frac{\kappa_{i,j}\overline{A}_{i,j}}{d_{i,j}}\left(h_j - h_i\right) + R_iV_i \tag{2}$$

where V_i is the volume of cell i and R_i accounts for the external stresses affecting the same cell. $\overline{A}_{i,j}$ is the cross-sectional area for the flow between cell i and any neighbouring cell j, $d_{i,j}$ is the distance between these two cells, and movement of the water between the cells is characterised by the hydraulic conductivity $\kappa_{i,j}$.

To further simplify the state-space, assumptions have to be considered regarding the cross-section $\overline{A}_{i,j}$. For a confined aquifer, $\overline{A}_{i,j}$ is constant and the state variable is described by a linear function since the boundaries for the aquifer do not evolve in time. Hence, the cross-section can be approximated by $\overline{A}_{i,j} = b_{i,j}E_{i,j}$ where $b_{i,j}$ is the thickness of the aquifer (bounded by confining layers above and below) and $E_{i,j}$ is the length of the cross-section (see Fig. 1). For unconfined aquifers, however, the thickness of the aquifer varies as a function of the water levels in the two cells i and j, thus, $\overline{A}_{i,j} = \overline{A}_{i,j}(h_i,h_j)$ and, hence, the dynamic description of the water level in cell i becomes nonlinear. In the following, a confined aquifer is considered, and further simplifications for the groundwater flow in the aquifer are from now on exclusively related to confined aquifers.

Defining the thickness of the confined aquifer as b, and by assuming the thick-

ness to be homogeneous, ($b_{i,j} = b_i = b$), the transmissivity of the aquifer can be defined as $T_{i,j} = \kappa_{i,j} b$, as well as the storage coefficient $S_i = S_{S,i} b$. Also, the base area of the cell can be obtained by $A_i = V_i / b$. Including $T_{i,j}$, S_i and A_i in Eq. (2), as well as replacing the constant $\overline{A}_{i,j}$ with $bE_{i,j}$ for a confined aquifer, the ith state in the state-space formulation is written

$$S_i A_i \frac{dh_i}{dt} = \sum_{j=1}^{J} \frac{T_{i,j} E_{i,j}}{d_{i,j}} (h_j - h_i) + W_i, \tag{3}$$

where W_i represents the sum of all sources and sinks in the ith cell (where sinks are presented with a negative sign).

The pumping rate Q_i in a well needs to be transferred to the cell, and consequently provides a description for the water drawdown in the aquifer as water is withdrawn from the well. This is taken care of by using the storage coefficient, which, in theory, accounts for this transfer (*Gupta*, 2008). If a water is discharged from a well, the affected cell needs to be recharged to keep the water balance. For a well field that only includes wells for absorbing water from the aquifer, the recharge must happen through the boundaries of the system. Here, this simply means that for a unbounded system the water level, far away from the well field, remains at a constant level as if no pumping occurred in the well field. Thus, the aquifer is constantly recharged though the barriers of the system, with a varying rate, depending on the difference between the water level h_i and H_0, which represents an upper boundary for h_i. The recharge of state i can be considered to be from above, through an area of the same size as the base area A_i.

Thus, the sum of sources and sinks can be formulated as

$$W_i = -S_i Q_i + R_i A_i (H_0 - h_i) \tag{4}$$

where, here, the external stress R_i corresponds to a leakage coefficient for the recharge of cell i. Replacing W_i in (3) with (4), the description of the state becomes

$$\begin{aligned} \frac{dh_i}{dt} &= \frac{1}{S_i A_i} \sum_{j=1}^{J} \frac{T_{i,j} E_{i,j}}{d_{i,j}} (h_j - h_i) + \frac{R_i}{S_i} (H_0 - h_i) - \frac{1}{A_i} Q_i \\ &= -\left(\frac{1}{S_i} \sum_{j=1}^{J} \frac{T_{i,j} E_{i,j}}{A_i d_{i,j}} - R_i \right) h_i + \frac{1}{S_i A_i} \sum_{j=1}^{J} \frac{T_{i,j} E_{i,j}}{d_{i,j}} h_j + \frac{R_i}{S_i} H_0 - \frac{1}{A_i} Q_i. \end{aligned} \tag{5}$$

Traditionally, the objective of solving the governing equation in well field applications is to provide a simulation for the water drawdown in the entire well field as the water heads in the wells respond to the water discharges from the

aquifer. This means that a large number of cells needs to be simulated. However, according to the model conceptualisation given above, predicting the future response in the wells only requires the states that are directly connected to the discharge wells, and with n wells connected in a well field the suggested state space model, at time t, becomes

$$d\begin{bmatrix} h_1(t) \\ \vdots \\ h_{n,t} \end{bmatrix} = \begin{bmatrix} -\frac{1}{S_1}\left(\frac{1}{A_1}\sum_{j=1}^{n}\frac{T_{1,j}E_{1,j}}{d_{1,j}} + R_1\right) & \cdots & \frac{T_{1,n}E_{1,n}}{S_1A_1d_{1,n}} \\ \vdots & \ddots & \vdots \\ \frac{T_{n,1}E_{n,1}}{S_nA_nd_{n,1}} & \cdots & -\frac{1}{S_n}\left(\frac{1}{A_n}\sum_{j=1}^{n}\frac{T_{n,j}E_{n,j}}{d_{n,j}} + R_n\right) \end{bmatrix} \begin{bmatrix} h_{1,t} \\ \vdots \\ h_{n,t} \end{bmatrix} dt + \begin{bmatrix} \frac{R_1}{S_1} & -\frac{1}{A_1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{R_n}{S_n} & 0 & \cdots & -\frac{1}{A_n} \end{bmatrix} \begin{bmatrix} H_0 \\ Q_{1,t} \\ \vdots \\ Q_{n,t}(t) \end{bmatrix} dt. \tag{6}$$

On a matrix form this can be written

$$d\boldsymbol{h}_t = [\boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{h}_t + \boldsymbol{B}(\boldsymbol{\theta})\boldsymbol{Q}_t]\,dt \tag{7}$$

where $\boldsymbol{\theta} \in \mathbb{R}^p$ is a vector of the parameters in the model and the matrices $\boldsymbol{A}(\boldsymbol{\theta}) \in \mathbb{R}^{n\times n}$ and $\boldsymbol{B}(\boldsymbol{\theta}) \in \mathbb{R}^{n\times m}$ describe the variation in the system dynamics for changes in the states and the input, respectively.

2.2 Grey box models

As previously stated, grey box models bridge the gap between physically-based models and statistical models. Hence, as the physical laws for models are typically formulated in continuous time and any observable data is in discrete time, the grey box model is merged in a continuous-discrete time description facilitated by the stochastic grey box model. The system description is in continuous time and to account for stochasticity in the system, it is formulated as a set of Stochastic Differential Equations (SDEs). This formulation of the system dynamics is herafter referred to as system equation of the state-space model. The states are partially observed in discrete time with a measurement noise, as described by the discrete time measurement equation.

The general expression for the grey box model, where the state variables are represented by the vector $\boldsymbol{h}_t \in \mathbb{R}^n$ and the variables for the model forcing by the vector $\boldsymbol{Q}_t \in \mathbb{R}^m$, is written

$$d\boldsymbol{h}_t = \boldsymbol{f}(\boldsymbol{h}_t, \boldsymbol{Q}_t, t; \boldsymbol{\theta})\,dt + \boldsymbol{\sigma}(\boldsymbol{Q}_t, t; \boldsymbol{\theta})\,d\boldsymbol{\omega}_t \qquad \text{(system equation)} \tag{8}$$

$$\boldsymbol{Y}_k = \boldsymbol{g}(\boldsymbol{h}_k, \boldsymbol{Q}_k, t_k; \boldsymbol{\theta}) + \boldsymbol{e}_k \qquad \text{(measurement equation)} \tag{9}$$

where $t \in \mathbb{R}_0$ is the time variable and k, for $k = 1, \ldots, K$, is the time instants for available observations; and $\boldsymbol{Y}_k \in \mathbb{R}^l$ is a vector of the measured output variables. The functions $\boldsymbol{f}(\cdot) \in \mathbb{R}^n$, $\boldsymbol{\sigma}(\cdot) \in \mathbb{R}^{n \times n}$ and $\boldsymbol{g}(\cdot) \in \mathbb{R}^l$ are, in general, nonlinear functions; $\{\boldsymbol{w}_t\}$ is a n-dimensional standard Wiener process; and $\{\boldsymbol{e}_k\}$ is a l-dimensional white noise process with $\boldsymbol{e}_k \in N(\boldsymbol{0}, \boldsymbol{S}(\boldsymbol{Q}_k, t_k, \boldsymbol{\theta}))$. The first term on the right-hand side of Eq. (8) is called the drift term, corresponding to the ODE for the dynamic structure of the system, while the second term is called the diffusion term. By only considering the ODEs in the system equation, the output error is usually autocorrelated. With the additional diffusion term to the ODEs in the system equation, a separation is provided between the model error and the measurement error, where the model noise contains the error for the approximated model description and deficient model forcing. This usually results in an uncorrelated measurement error, and a model structure that is adjusted to the output measurements.

A simplified version of the grey box approach is obtained by considering the drift term in the system equation to be linear and time-invariant. Also, the function $\boldsymbol{g}(\cdot)$ is linearly related to both $\boldsymbol{h}_t$ and $\boldsymbol{Q}_t$, hence the grey box model in Eq's. (8) and (9) can be written:

$$d\boldsymbol{h}_t = [\boldsymbol{A}(\boldsymbol{\theta})\,\boldsymbol{h}_t + \boldsymbol{B}(\boldsymbol{\theta})\,\boldsymbol{Q}_t]\,dt + \boldsymbol{\sigma}(\boldsymbol{Q}_t, t, \boldsymbol{\theta})\,d\boldsymbol{\omega}_t \tag{10}$$

$$\boldsymbol{Y}_k = \boldsymbol{C}(\boldsymbol{\theta})\,\boldsymbol{h}_k + \boldsymbol{D}(\boldsymbol{\theta})\,\boldsymbol{Q}_k + \boldsymbol{e}_k \tag{11}$$

where the matrices $C(\boldsymbol{\theta}) \in \mathbb{R}^{l \times n}$ and $D(\boldsymbol{\theta}) \in \mathbb{R}^{l \times m}$ relate, respectively, the state variables and the input data to the measured output. The drift term in the system equation (10) corresponds to the state-space formulation for the well field model in (7). Thus, the system equation (10) is formulated by the ODE in (6) plus an additional diffusion term. However, the diffusion term $\boldsymbol{\sigma}(\cdot)$ in (10) is considered to be time-varying and can also be expressed by a nonlinear function of the input variables.

2.3 Parameter and state estimation

For estimating the model parameters the Maximum Likelihood (ML) method is used. The likelihood function can be evaluated by applying Kalman Filtering techniques for continuous-discrete time state space models (*Kristensen et al.*, 2004, *Jazwinski*, 2007). For a sequence of observations, $\mathcal{Y}_k = [\boldsymbol{Y}_k, \cdots, \boldsymbol{Y}_0]$, the likelihood function L is calculated as a product of one-step ahead predictive conditional densities for all the observations (*Madsen*, 2008). Hence, an optimum parameter set $\boldsymbol{\theta}$ is seeked, such that a maximum value is found for the likelihood function

$$L(\boldsymbol{\theta}; \mathcal{Y}_k) = \left(\prod_s^k p(\boldsymbol{Y}_s | \mathcal{Y}_{s-1}, \boldsymbol{\theta}) \right) p(\boldsymbol{Y}_0 | \boldsymbol{\theta}). \tag{12}$$

For the linear state space model as specified by (10) and (11), all the densities are Gaussian. Then, for the grey box model, the objective is to maximise a likelihood function, containing a product of one-step conditional Gaussian densities. The one-step ahead prediction mean $\hat{\boldsymbol{Y}}_{k|k-1} = E\{\boldsymbol{Y}_k|\mathcal{Y}_{k-1},\boldsymbol{\theta}\}$ and the covariance $\boldsymbol{R}_{k|k-1} = V\{\boldsymbol{Y}_k|\mathcal{Y}_{k-1},\boldsymbol{\theta}\}$ characterise the Gaussian density and, hence, the likelihood function in (12) can be rewritten

$$L(\boldsymbol{\theta};\mathcal{Y}_k) = \left(\prod_{s=1}^{k} \frac{\exp\left(-\frac{1}{2}\boldsymbol{\epsilon}_k^\top \boldsymbol{R}_{s|s-1}^{-1}\boldsymbol{\epsilon}_s\right)}{\sqrt{\det(\boldsymbol{R}_{s|s-1})}\left(\sqrt{2\pi}\right)^l} \right) p(\boldsymbol{Y}_0|\boldsymbol{\theta})$$

where $\boldsymbol{\epsilon}_k = \boldsymbol{Y}_k - \hat{\boldsymbol{Y}}_{k|k-1}$ and the conditional mean and covariance are recursively computed using the Kalman filter Ṫhus, by conditioning on $\boldsymbol{Y}_0$ the parameter estmates can be obtained by maximising with respect to the parameter $\boldsymbol{\theta}$, i.e.

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \left(\ln\left(L(\boldsymbol{\theta};\mathcal{Y}_K|\boldsymbol{Y}_0)\right)\right).$$

The states are also included in the updating procedure of the Kalman filter, such that an estimation of the unobserved states is provided, as well as gaps in the data series for the observable states can be bridged by the filtering method (*Kristensen and Madsen*, 2003, *Kristensen et al.*, 2004, *Jonsdottir et al.*, 2006).

3 The test case and available data

The case study considers the well fields connected to the Søndersø Waterworks, which is located in the Northern part of Zealand, Denmark. A map of the Søndersø lake and its surroundings is shown in Figure 2, including all operating wells connected to the waterworks. The waterworks consist of three subgroups of wells, where two are located close to the lake and the third one is along the river Tibberup, which diverts water from Søndersø lake. The whole waterworks delivers 8 million m^3 of water each year, abstracted from aquifers in the well field by 21 abstraction wells. However, observations from each well are only available from the wells at Søndersø East and West well fields, and not for the wells along Tibberup river, and since the objective of the model is to interpret the water level in the individual well, the 10 wells along the river Tibberup are excluded in the following.

The available data consist of measurements from 11 wells, where a cluster of 8 wells is on the North and East shore of lake Søndersø and 3 wells are located South of the lake. For the study, the requirement is that the wells should all

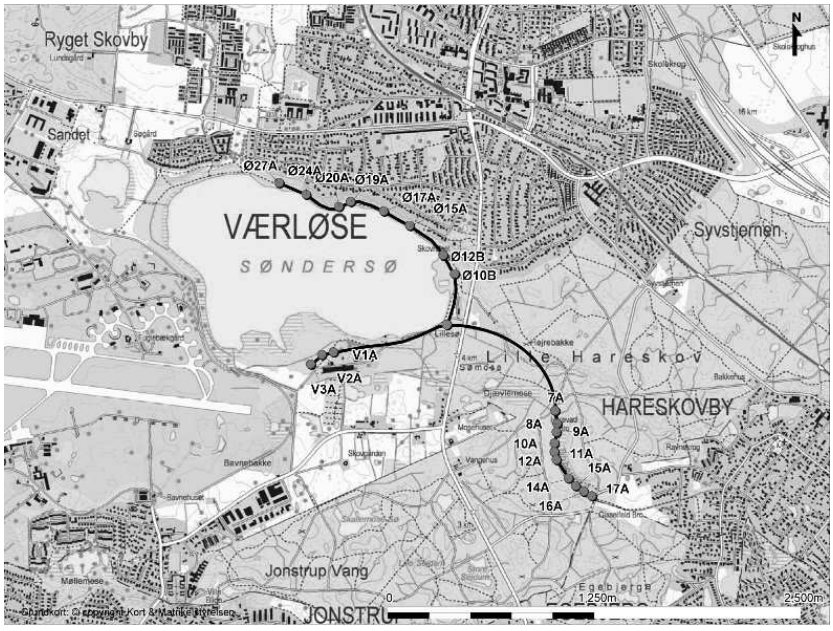

Figure 2: Map of the Søndersø waterworks and surroundings. The 21 wells, attached to the waterworks, are marked with dots; the East wells with "Ø", the West wells with "V", and the wells along Tibberup river are marked "7A"-"17A".

be pumping from the same aquifer. One of the wells is not pumping from the same aquifer as the remaining wells. In Figure 2 this well is marked Ø20A, and it is excluded in the following.

The 10 wells all abstract water from the limestone aquifer 20-50 meter below surface. The aquifer is neither in hydraulic contact with the lake nor the Tibberup stream. For each well, the discharge rate for each minute is registered, along with corresponding water drawdown in the well. The data used for estimation spans the period from December 8th, 2008, to January 26th, 2009, and with the one minute resolution the period consist of 120,260 time instants. The available data for the limestone aquifer wells is displayed in Figure 3. The negative correlation between the head in the wells and the corresponding discharge is clearly seen in the figure. The figure also shows the correlation between water levels in different wells, as changes in a particular well have influence on the other wells, penetrating the same aquifer, and the strength of this correlation appears to be inversely related to the distance between wells. This corresponds to the model obtained in Eq. (6) for the water levels in the wells. Moreover, as seen in (6) the grey box model requires measurements for the dis-

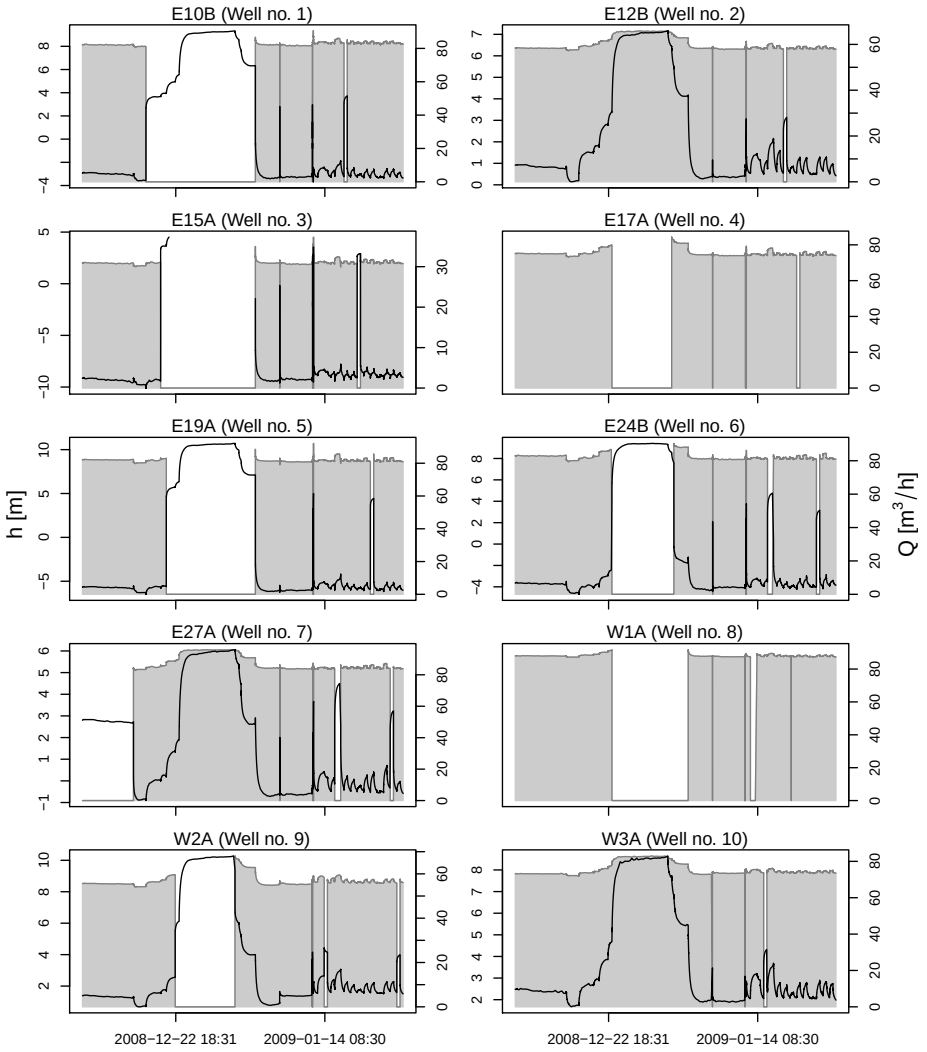

Figure 3: Available data for modelling the limestone aquifer. In each panel the black line is the water drawdown in the well (axis to the left) and a corresponding discharge flow rate is presented by the grey area (axis to the right).

tance $d_{i,j}$ between wells i and j, but for the Søndersø well field these measures are available and are, therefore, not required in the parameter estimation.

One important observation from Figure 3 is the missing data in the time series used for the modelling approach. No water level measurements are available for Well 4 and Well 8. However, these wells cannot be neglected from the model since the wells are abstracting water from the same aquifer and, consequently, their water discharges have significant influence on the water levels in all other wells. Therefore, the water level of Wells 4 and 8 is not considered in the vector of measured variables for the grey box model. Furthermore, the water level measurements for Well 3 are partly missing, but the missing time series is estimated by applying Kalman filter updating of the state for Well 3 in the system equation.

4 Results

For illustrating the benefits of the grey box approach, three models are proposed and compared.

4.1 Model 1: lumped parameter model

The first model in the study is a lumped parameter model of the stochastic groundwater model, introduced in Section 2. A lumped parametrisation provides a more condensed description of the system, where both the model and the parameters are identifiable from data.

The ith state in the lumped parameter model (hereafter called Model 1) is similar to the one given in (5), but with few assumptions regarding the parameters. First, it is difficult to distinguish between the transmissivity $T_{i,j}$ and corresponding cross-sectional length $E_{i,j}$. Therefore, instead the transmissivity for each cross-sectional unit, i.e. the product $\overline{T}_{i,j} = T_{i,j}E_{i,j}$, is estimated. Furthermore, the aquifer is assumed to be homogeneous and isotropic, and since all wells pump from the same aquifer the transmissivity $\overline{T}$ is considered to be the same between any two wells. Second, the storage coefficients S is a parameter that characterises the aquifer and is assumed to be the same for all states in the system equation. Third, semipervious confining layers are the main source of recharge of cells, and for recharging the cells this source is assumed to have the same charateristics everywhere in the well field, implying that the leakage coefficient R can be considered as a constant term. Fourth, to further reduce the number of parameters the area A, the base area of a cell that includes a single

well, is assumed to be the same for all cells. This is a rather harsh assumption, since the boundaries of each cell should correspond to no-flow lines, but such boundaries usually require continuous adjustment due to changes in the aquifer flow (*Narasimhan and Witherspoon*, 1976, *Rozos and Koutsoyiannis*, 2010). However, due to parsimonious reasons this simplification is adopted as well. Hence, assumptions are listed as follows:

1. $\overline{T} = T_{i,j}E_{i,j}$, for $i,j = 1,\ldots,10; i \neq j$,
2. $S = S_1 = S_2 = \cdots = S_{10}$,
3. $R = R_1 = R_2 = \cdots = R_{10}$,
4. $A = A_1 = A_2 = \cdots = A_{10}$,

and by including these in Eq. (5), the ith state in the system equation becomes

$$dh_i = \left[-\left(\frac{9\overline{T}}{SA} \sum_{j=1 j\neq i}^{n} \frac{1}{d_{i,j}} + \frac{R}{S} \right) h_i + \frac{\overline{T}}{SA} \sum_{j=1 j\neq i}^{n} \frac{h_j}{d_{i,j}} + \frac{RH_0}{S} - \frac{Q_i}{A} \right] dt + \sigma_i d\omega_i, \tag{13}$$

where the diffusion parameter σ_i is a constant term, estimated along with the model parameters by the ML method. The measurement equation for Model 1 is considered to be directly observable for wells with available data for the water level. Of the 10 wells included in the system equation, only 8 are observed (see Fig. 3), indicating that the measurement equation for well i becomes

$$H_i = h_i + e_i \tag{14}$$

where H_i represents the measured water head in well i. On a matrix form the grey box model can then be written

$$d\boldsymbol{h}_t = [\boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{h}_t + \boldsymbol{B}(\boldsymbol{\theta})\boldsymbol{Q}_t]\,dt + \boldsymbol{\sigma} d\boldsymbol{\omega}_t \tag{15}$$

$$\boldsymbol{H}_k = \boldsymbol{C}(\boldsymbol{\theta})\boldsymbol{h}_k + \boldsymbol{e}_k. \tag{16}$$

This expression for the stochastic well field model is a subset of the more general version in Eq's. (10) and (11), where the diffusion $\boldsymbol{\sigma}(\cdot) = \boldsymbol{\sigma}$ is a constant diagonal matrix and $\boldsymbol{D}(\boldsymbol{\theta}) = \boldsymbol{0}$ since the input variables are not considered to have direct influence on the observed output.

The parameter estimates for Model 1 is shown in Table 1. The estimates for the area A and H_0 are seemingly physically interpretable. If A is assumed to be a circle, with a pumping well at origin, the radius would be approximately 23m. H_0 is a little higher than the maximum water level in the wells, implying that water is constantly being recharged in the well field. For a single length

unit of the cross-section ($E_{i,j} = 1$) the estimated transmissivity in the aquifer $\overline{T}$ can by interpreted physically, and compared with pumping test results from the Søndersø well field, where the average transmissivity was 30.5m^2/h, the estimate is seemingly adequate. Both the storage coefficient S and the leakage R are also physically meaningful, but the estimate of S is too high (the storage

Table 1: Estimation Results for Models 1-3. Bold face numbers refer to parameter estimates, and below each estimate is a corresponding standard deviation (parenthesis).

$\hat{\theta}$	Units	Model 1	Model 2	Model 3
		Drift parameters		
S	[-]	$\mathbf{2.7}\cdot10^{-2}$	$\mathbf{4.3}\cdot10^{-3}$	$\mathbf{4.7}\cdot10^{-3}$
		$(0.5\cdot10^{-2})$	$(0.2\cdot10^{-3})$	$(0.2\cdot10^{-3})$
$\overline{T}$	[m·m^2/h]	**26.16**	**13.41**	**14.14**
		(3.80)	(1.28)	(0.98)
T_{E1}	[m^2/h]		**11.46**	**11.72**
			(0.10)	(0.07)
T_{E2}	[m^2/h]		**28.85**	**29.35**
			(0.28)	(0.19)
T_3	[m^2/h]		**1.620**	**1.711**
			(0.028)	(0.029)
T_W	[m^2/h]		**103.0**	**104.8**
			(5.5)	(5.0)
R	[h^{-1}]	$\mathbf{1.3}\cdot10^{-4}$	$\mathbf{3.1}\cdot10^{-5}$	$\mathbf{3.0}\cdot10^{-5}$
		$(0.5\cdot10^{-4})$	$(0.3\cdot10^{-5})$	$(0.1\cdot10^{-5})$
A	[m^2]	**1598.0**	**1411.5**	**1731.9**
		(295.6)	(128.9)	(88.2)
A_3	[m^2]		**4.13**	**5.36**
			(0.35)	(0.49)
H_0	[m]	**15.28**	**11.42**	**10.96**
		(1.88)	(0.51)	(0.29)
		Diffusion parameters		
σ_1		**0.451**	**0.122**	**0.123**
		(0.010)	(0.002)	(0.002)
σ_2		**0.130**	**0.092**	**0.094**
		(0.003)	(0.002)	(0.002)
σ_3		**1.254**	**0.844**	**0.830**
		(0.026)	(0.023)	(0.023)
σ_5		**0.729**	**0.462**	**0.472**
		(0.014)	(0.010)	(0.010)
σ_6		**0.611**	**0.188**	**0.192**
		(0.012)	(0.005)	(0.005)
σ_7		**0.246**	**0.121**	**0.125**
		(0.005)	(0.003)	(0.003)
σ_9		**0.232**	**0.122**	**0.126**
		(0.047)	(0.003)	(0.003)
σ_{10}		**0.117**	**0.099**	**0.101**
		(0.003)	(0.002)	(0.002)

coefficient from the previously mentioned pumping test was $2.64 \cdot 10^{-4}$) and, consequently, the estimated leakage coefficient is a little higher than expected.

Regarding the estimated diffusion parameters, the estimates seem to be rather high for some of the wells in the system equation (σ_1, σ_3, σ_5 and σ_6). A large estimate of the diffusion parameters propagates through the system equation and is observed as a large variance of the predicted water levels. This means that the prediction interval for the water heads is proportional to the estimated diffusion in the system equation. The prediction intervals for the available water heads are displayed in Figure 4 (enclosed by the black dash lines). The plots show that the intervals are too large for any application of the water head predictions in the wells, especially if a sequences of observations of a water head is missing in the time series (see the infeasible estimates for Well 3 in Fig. 4).

It is obvious from both the estimated parameters and the evaluated prediction interval, that Model 1 is not able to provide reliable water head predictions with a reasonable assessment of the embedded uncertainty. Thus, the stochastic well field model requires an improvement.

4.2 Model 2: Including well equation

Assuming the states to be the measured water levels in the measurement equation (14) is inconsistent with the approach for the forcing in the system equation (13), since the system is assumed to be the aquifer where the discharge variable is transferred to the aquifer with the storage coefficient S. This has to be corrected in the measurement equation because the system is not representing the head in the well but the head in the cell that includes the well. Hence, the observation equation for well i needs to contain a function for the well losses that are present when water is pumped from well i. Then the measurement equation (14) becomes

$$H_i = h_i - D_i Q_i + e_i \tag{17}$$

where D_i is a function for the loss in the water head between the well and the cell. This head difference can be described with a well-loss functions, consisting of a linear term for the aquifer loss and nonlinear term for the well loss that, respectively, correspond to the head loss through the well screening and the head loss from the perimeter of the cell to the well screening (*Hansen et al.*, 2011). In this study however, the function is considered to be time-invariant and linearly related with the pumping rate, since the heads in the wells appear to approach a steady-state condition. Thus, the well loss and the aquifer loss

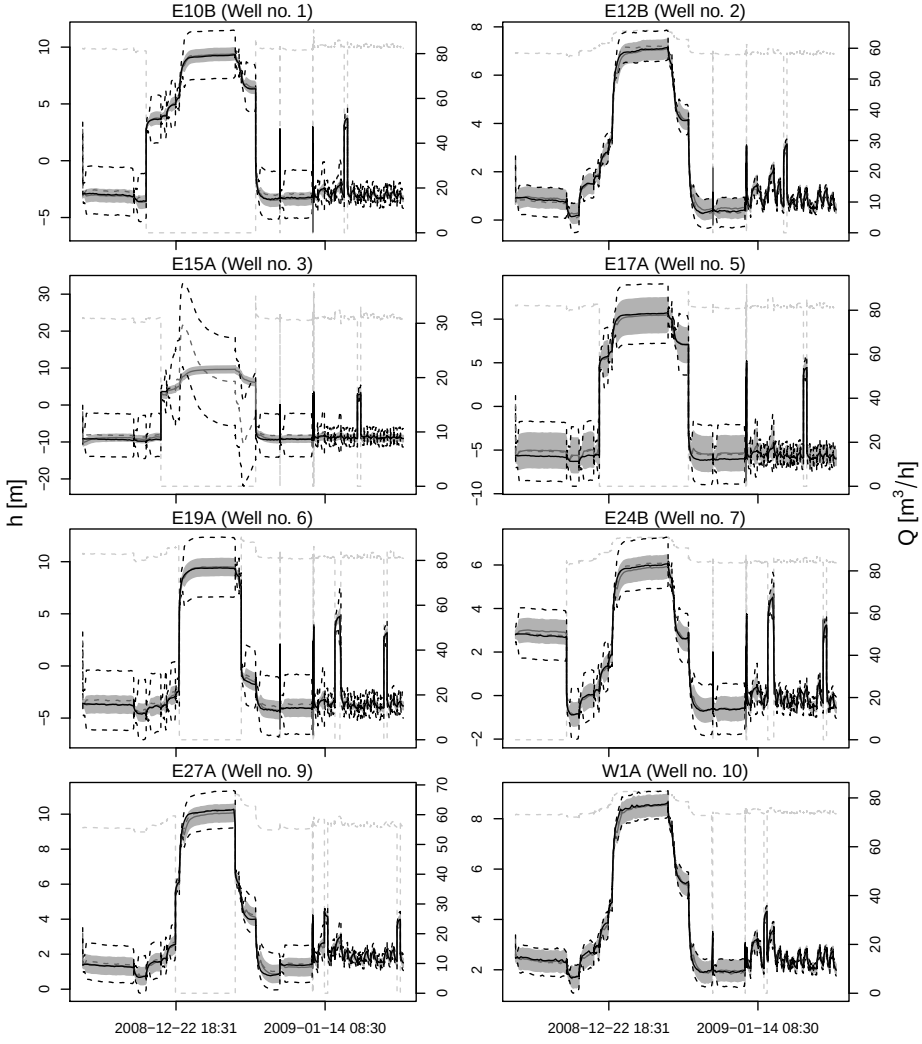

Figure 4: Comparison for 95% prediction intervals between Model 1 (black dashed lines) and Model 2 (grey shaded area). The observed values for the water level are connected by the black solid line and the predictions for Model 1 and Model 2 are shown as dark grey dash and solid lines, respectively. The light grey dash line is the pumping rate for the corresponding well (read from the right label in each plot).

as represented with an aggregated loss function:

$$D_i = \frac{\ln\left(\frac{r_i}{r_{w,i}}\right)}{2\pi T_i} \tag{18}$$

where T_i is the transmissivity for the flow from the aquifer to the well, $r_{w,i}$ is the radius of the well, and r_i is the radius of the part of the aquifer that is affected by the discharge. The loss function (18) is uniquely defined for each well. If the estimated area A in the model is approximated to be a circle, r_i can be replaced in the loss function, and the measurement equation for well i is written

$$H_i = h_i - \frac{\log\left(\frac{\sqrt{A_i/\pi}}{r_{w,i}}\right)}{2\pi T_i} Q_i + e_i. \tag{19}$$

Thus, the updated grey box model consist of the system equation in (13), but the observation equation is now identical to the one represented in (11) where $\boldsymbol{D}(\boldsymbol{\theta})$ is a matrix with the well losses D_i on its diagonal. This model will be referred to as Model 2 in this study. The transmissivity T_i refers to the transmissivity from distance r_i to the well, including the flow through the well filter. This transmissivity should be distinguished from $\overline{T}$, but to prevent a possible overparametrisation by assigning a new parameter to each well the geographical location is considered to find a reduced number of transmissivity parameters for the estimation. Two parameters of transmissivity are introduced for the wells on the east side, marked T_{E1} for wells 1, 5 and 6, and T_{E2} for wells 2 and 7; and one for the wells on west side: T_W.

Furthermore, the data in Figure 3 reveals that the variation in the water drawdown for Well 3 is one of the largest (similar to the variation for Well 1), but for a much lower pumping rate. This indicates that the characteristics for Well 3 are not the same as for the other wells and an adjustment is needed in the parametrisation with focus on Well 3. The increasing flow to the well is mainly caused by the transmissivity T and the area A, and from the data it becomes clear that these two parameters have to be estimated for Well 3 separately from the other wells. Hence, we introduce the parameters T_3 and A_3 in the estimation to characterise the water drawdown in Well 3.

The estimation results for Model 2 are shown in Table 1, and compared to the results for Model 1 some great improvements are achieved. By considering the measurement equation as a well equation for Model 2 the recharge has been reduced since both H_0 and R are estimated much smaller than in Model 1. Also, the storage coefficient S is corrected towards a lower value. The lumped parameter estimate for the transmissivity in the aquifer $\overline{T}$ is now only a half of the previous estimate. By distinguishing between the transmissivities in well screening and the lumped transmissivity in the aquifer, $\overline{T}$ becomes a more

unique property of the aquifer. The result for the estimated transmissivities T_{E1}, T_{E2}, T_3 and T_W are clearly distinct, since the estimates are not remotely close. The deviation between the east side transmissivities is the smallest, but far from being considered the same. The small value for T_3 is as expected, because changes in the water head in Well 3 are (almost) exclusively related the discharge in the Well 3. In contrast, the wells on the west side are highly depending on the discharges from some of the other wells, which is reflected in the high estimate for T_W. The area A is similar for both models. By particularly estimate the base area for Well 3 a fairly small value is obtained for A_3, which corresponds to the large variation in the water head for such a low pumping rate. For the diffusion parameters, a prominent decrease in the parameter estimates is attained in favour of Model 2. Especially for uncertainty parameters that were estimated to be unrealistically high in Model 1 are now significantly improved, i.e. σ_1, σ_3, σ_5 and σ_6 are reduced by approximately $30 - 75\%$.

Changes in the one-step prediction and the improvement for the prediction intervals can be visualised for each well in Figure 4. It is difficult to observe the difference between the two predictions, but that is mainly due to the fact that these two predictions are almost identical, as well as identical to the black solid line, representing the observations of the water level. It can be discussed which of these two is the better candidate for the prediction, but the main conclusion is that they give almost the same predictions and the performance improvement is determined by the prediction intervals. By comparing the region embedded within the black dash lines (Model 1) and the grey shaded area (Model 2), a large decrease of the prediction intervals is detected, where the reduction for each well is in accordance with the reduction of the related diffusion parameter from Table 1. This correction for the intervals is mainly due to the inclusion of a well function in the measurement equation in the grey box model approach. Also, the missing data in the water level of Well 3 are now estimated sufficiently by including the well function in the measurement equation of the grey box model. Model 1 is not able to provide sufficient estimates for the states in the model, but with Model 2 the states and their uncertainties are adequately assessed.

4.3 Model 3: Formulating the diffusion

Although the physical parameters for Model 2 are seen to be reasonably estimated, and it appears from Figure 4 that the one-step prediction is able to fit the observations, the main concern regarding Model 2 is however the estimation of the prediction interval. The model improvement from Model 1 to Model 2 shows a large reduction in the estimated interval, but still it is too wide when the water heads approach a steady-state condition. The challenge

here would be to give a "right" interpretation of the prediction interval as time between changes in the pumping is prolonged. In the model structure the diffusion requires a functional form, such that the state uncertainties are rapidly increased as pumps are either switched on or off, but subsequently decreased until a steady-state is reached and the uncertainty is at its minimum.

With the on-off setting for the pumps, a feasible function for the diffusion would be an Impulse Response Function (IRF), inspired by the physical parameters already included in the model structure. For a pumping well, one such physically-based IRF is related to the Hantush formula, and describes a penetrating well in an unbounded aquifer. The Hantush formula can be defined by

$$M_i = -\frac{1}{4\pi T_{i,j}\tau_{i,t}} \exp\left(-\frac{d_{i,j}^2 S_i}{4T_{i,j}\tau_{j,t}} - \frac{R_i \tau_{i,t}}{S_i}\right) \tag{20}$$

where all parameters have been defined in Section 2, and $\tau_{i,t} = t - t_{i,0}$, where $t_{i,0}$ corresponds to the preceding time instant pump i was turned on or off. The first term in the exponential function expresses the delayed influence on well i when pumping from well j at distance $d_{i,j}$. However, the data shows that these influences are vanishing when compared with the instantaneous changes in well i for the same pumping at well j. The effect of the changes in the discharge are observed in almost all the other wells instantaneously. Thus, the first term in the exponential function can be disregarded in the diffusion term of the grey box model, and only one time variable τ_t is used as an input to the grey box model, accounting for all on-off shifts in the discharges for all wells in the study. Further, due to parsimonious reasons, the time-varying term, multiplied to the exponential function in Eq. (20), is considered to be a constant and accounts for the span of the uncertainty when τ_t is initiated. With this simplified Hantush formula, the diffusion becomes

$$\sigma_{i,i}(\boldsymbol{Q}_t, t; \boldsymbol{\theta}) = \sigma_i \exp\left(-\frac{R}{S}\tau_t\right). \tag{21}$$

By including this formula for all 10 states in the system equation (10), the diffusion terms are now input dependent and nonlinear, in addition to the linear and time-invariant drift term in the model in (6). Combined with the same measurement equation as for Model 2, the third model proposal for this study is obtained (hereafter referred to as Model 3).

It is not surprising that the estimated parameters in Model 2 and Model 3 (see Table 1) are quite alike, since the physical system is identical. The physical estimates that are expected to be adjusted when moving from Model 2 to Model 3 are the ones used in the diffusion formulation; S and R, but neither S nor R are significantly affected by this modification in the model structure. More

interesting is to see how the uncertainty has been modified, and a comparison between the two models is shown in Figure 5 where, now, the black dashed lines are the limits for the prediction interval of Model 2 and the grey region is the prediction interval for Model 3. In each panel in the figure the one-step prediction is plotted for the two models (dark grey lines, dash and solid for Models 2 and 3, respectively), and since the physical structure of the models is the same, and the parameter estimates very similar, the predictions also become very alike (for most of the plots in Fig. 5 it is difficult to distinguish between the predictions since they almost overlap). The figure shows very clearly how the uncertainty of the water level in each well is reduced as the time between changes in the pumping rates increases. The exponential decay is especially noticable for the wells where the uncertainty is rather high, as the reduction is more rapid (Wells 5 and 7). Also, the decreasing prediction interval seems to adjust to the accuracy of a corresponding prediction, i.e. decreasing noise between the prediction and the measurement results in a faster reduction of the prediction interval. This is a consequence of the estimated diffusion parameters in Model 3, which represent the amplitude of the diffusion. By further extending the diffusion term, the prediction interval could be improved, but at the expense of the parametrisation of the diffusion matrix $\boldsymbol{\sigma}(\cdot)$ in (10). Further extensions for the diffusion term in the grey box model are not considered in this paper.

The improved predition intervals for the water levels can be clearly seen; first in Figure 4 when comparing Models 1 and 2, and then again in Figure 5 for comparison of Model 2 and 3. However, only considering results seen from figures are not enough to evaluate the adoption of the prediction interval in the model. Thus, a quantification of the prediction performance is required.

4.4 Evaluating the prediction intervals

The classical way of evaluating predictions is by quantifying the deviation between the predictions and corresponding measurements. Many methods have been proposed to find the "right" measure of model performance; e.g. the mean square error (MAE), the root mean square error (RMSE), the mean average error (MAE) and the standard deviation error (SDE) (see *Madsen et al.*, 2005). The most popular criterion though for evaluating hydrological models is the Nash-Sutcliffe coefficient (*Nash and Sutcliffe*, 1970). For the grey box model, or models that apply the maximum likelihood method for estimating the unknown parameters, it is straight forward to use the likelihood, given in Section 2.3, since the optimal parameter set is simply obtained when the likelihood is maximised. All these methods (and many more) provide a single number for the quantification and are good representitives for the evaluation

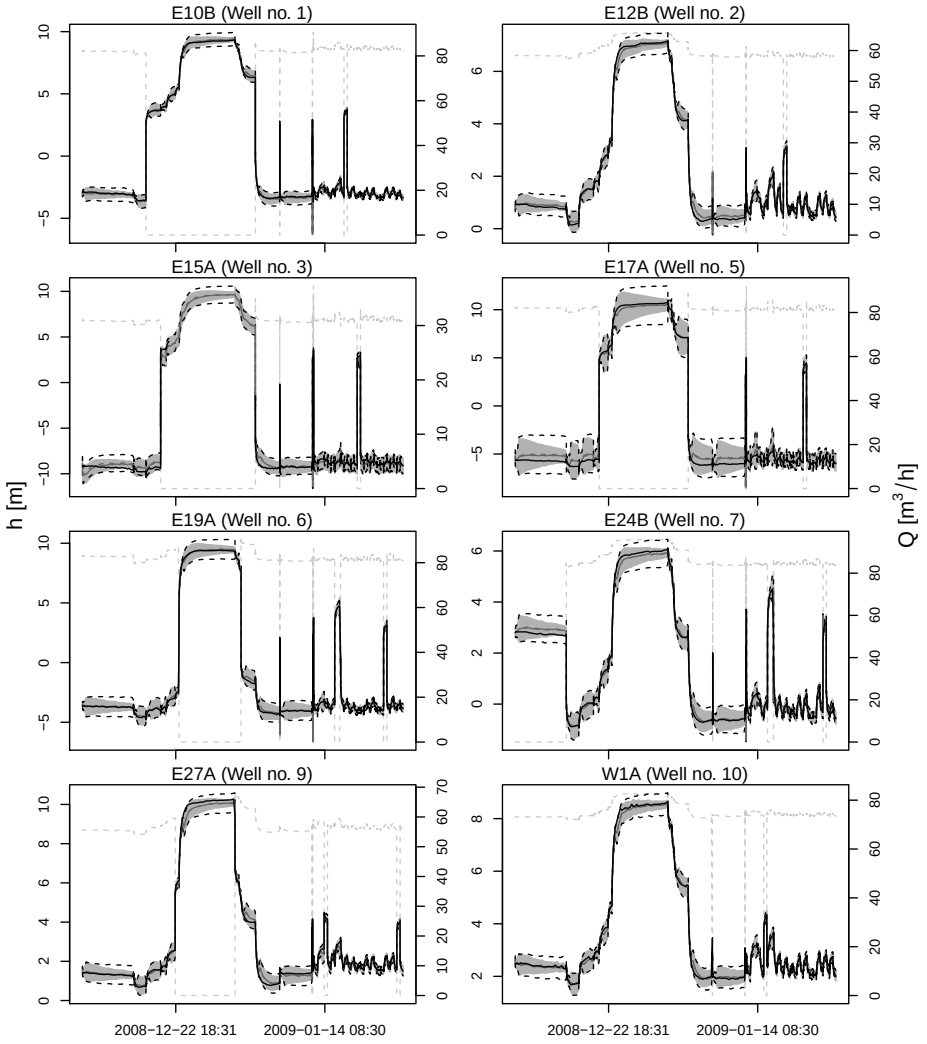

Figure 5: Comparison for prediction intervals between Model 2 (black dashed lines) and Model 3 (grey shaded area). The observed values for the water level are connected by the black solid line and the predictions for Model 2 and Model 3 are shown as dark grey dash and solid lines, respectively (it is difficult to see the difference between the predictions because the lines overlap). The light grey dash line is the pumping rate for the corresponding well (read from the right label in each plot).

of so-called point predictions. Here, however, the aim is to quantify the prediction intervals by a single number, similar to the methods described above for the point prediction.

The method used in the following compares the prediction 95% quantiles with the available output data. This can be inspected visually by comparing the prediction limits in Figures 4 and 5 with the included measurements, but to quantify the performance of the predition interval a scoring criterion is required that merges all relavant information from the interval into a single number. For a model to be even considered as an adequate candidate, a reliability requirement has to be fulfilled, i.e. the proportion of the observed data inside the estimated prediction interval should be the same as the predetermined coverage of data (here, the coverage is 95%). For any deviation between the coverage and the observed proportion the reliability of the model is lost and a bias is detected between the model and the observations. Further, the prediction intervals must have narrow regions, since too large intervals will reduce the accuracy of the predicted outcome and create unwanted scenarios for decision makers. The size of the prediction interval is often referred to as sharpness (*Gneiting et al.*, 2007).

To quantify the performance of the model both the reliability and the sharpness have to be accounted for in the evaluation. The sharpness does not provide any information regarding the observed values since only the upper and lower limits of the prediction interval are considered. In contrast, the reliability provides information about the observed data, but only as an indicator of hits and misses of the observed values within the prediction interval and fails to inform about the size of the region. However, what is not accounted for in these two performance measures is the deviation between an observation that fails to be within the prediction interval and the interval itself. Hence, this additional feature should be combined with both reliability and sharpness in a single number (called the skill score) for the performance evaluation. Thus, a proper skill score criterion is applied for the prediction interval (restricted by an upper and a lower quantile). For an upper quantile u and a lower quantile l, for a coverage β, the interval score (*Gneiting and Raftery*, 2007) is calculated

$$Sc_\beta(l,u;Y) = (u-l) + \frac{2}{\beta}(l-Y)\mathbb{I}\{Y<l\} + \frac{2}{\beta}(Y-u)\mathbb{I}\{Y>u\} \qquad (22)$$

where $\mathbb{I}\{\cdot\}$ is an indicator variable equal to one if the statement inside the brackets is fulfilled, but zero otherwise. This is an attractive approach because the sharpness is included directly $(u-l)$ and observations that are outside the prediction interval are penalised in accordance with the span of this departure away from the closest interval limit. Thus, the best model candidate for the 1-step prediction interval is the model with smallest score value.

The skill score criterion (22) for all the available water level data, evaluated for all three models, are listed in Table 2. The table also lists the proportion of observations that were detected ouside the estimated prediction interval (α). It is not surprising that the skill score for Model 1 is significantly larger than the score for the other two models for most wells, since the size of the predicition interval, or the sharpness, is much larger. This can best be detected by observing wells 2 and 10 in Figure 4 where the limits of the prediction region for Model 1 approaches the interval for Model 2, and correspondingly the skill score values for the models are closer to each other. Even though, for all observed wells, the proportion of misses for Model 1 (α_1 in Table 2) is smaller than the corresponding proportion of misses of the other models, it is not apparent in the resulting score values when penalisation in the score related to the size of the interval ($u - l$) is taken into account. This rules out Model 1 as the best model candidate for the prediction interval of the water drawdown in the observed wells.

The score results for Model 2 and Model 3 are quite similar where, for almost all wells, Model 3 is performing a little better. It is only for Well 3 where Model 2 is outperforming Model 3. The reduction of the prediction interval in Model 3 is due to the exponential decay as the duration between on-off regulations for the pumps is increased. Proportion of misses for these two models are almost identical for all wells, implying that the difference in the skill score is related to the sharpness of the models. Hence, formulating the diffusion term in the grey box model as a function of the input signal (Model 3) results in more narrow prediction intervals and an improvement in the model performance.

Table 2: Performance evaluation by using skill score criterion. A skill score is calculated for each well for comparison of the models, where the best model candidate has the lowest score value. Also in the table, the proportion of observations outside the predition interval (α) is shown.

		Well 1	Well 2	Well 3	Well 5	Well 6	Well 7	Well 9	Well 10
Model 1	$Sc_\beta^{(1)}$	3918	1105	8831	5842	5341	2139	1828	963
	α_1	0.013	0.024	0.016	0.019	0.018	0.019	0.020	0.047
Model 2	$Sc_\beta^{(2)}$	1180	813	4442	3746	1918	1056	969	815
	α_2	0.026	0.044	0.025	0.020	0.028	0.044	0.039	0.060
Model 3	$Sc_\beta^{(3)}$	1132	782	4479	3590	1853	1016	955	789
	α_3	0.026	0.043	0.025	0.020	0.028	0.043	0.039	0.060

5 Conclusions

The study has demonstrated how the grey box model can by applied for modelling water heads in wells, which pump water from a confined aquifer in a well field. The grey box model consists of a system equation and a measurement equation, and by replacing the governing equations with a set of stochastic differential equations in the system equation the model structure for the groundwater flow was simplified, but without losing the essential physical interpretation of the system. The parameters in the simplified model were sufficiently estimated, and from the estimation results both the drift term and the diffusion term of the system equations were further developed to obtain an improved stochastic model structure for the well field. Three grey box models were estimated, which all differed in the model structure and were gradually improved to cope with both the dynamics in the aquifer and the uncertainties embedded in the structure of the system. The first model was a lumped parameter model and showed the feasibility of the model approach. For the second model a well function was included, which led to improved parameter estimates and, consequently, a refinement of the system dynamics. The third model had a diffusion term that varied with changes in the on-off setting of the pumping rates, and showed further improvement of the second model, especially for the uncertainty of the one-step prediction intervals. Finally, the performances of the models were quantified by using the skill score criterion, which verified these findings for the three models. The grey box model provides a simple structure for the dynamics of the system where both model and its parameters are identifiable from data, and, thus, it is believed that such models will be well-suited for online forecasts and control of well fields.

Acknowledgements

This work was funded by the Danish Strategic Research Council, Sustainable Energy and Environment Programme, as part of the Well Field Optimisation project (http://wellfield.dhigroup.com/)

References

Adams RA (1999) *Calculus: a complete course*, 4th Edition. Addison Wesley Longman Ltd., Canada, Ch. 16, pp. 946–953.

Ahlfeld DP, Baro-Montes G (2008) Solving unconfined groundwater flow management problems with successive linear programming. *Journal of Water Resources Planning and Management* **134** (5):404–412.

Anderson MP, Woessner WW (2002) *Applied groundwater modeling: simulation of flow and advective transport*. Academic Press, California, USA.

Bacher P, Madsen H (2011) Identifying suitable models for the heat dynamics of buildings. *Energy and Buildings* **43** (7):1511–1522.

Bauser G, Henrik-Franssen HJ, Kaiser HP, Kuhlmann U, Stauffer F, Kinzelbach W (2010) Realtime management of an urban groundwater well field threatned by pollution. *Environmental science & technology* **44** (17):6802–6807.

Bear J (1979) *Hydraulics of Groundwater*. McGraw-Hill, New York,USA.

Beven K (1996) A discussion of distributed hydrological modelling. In: *Disributed Hydrological Modelling*. Abbott MB, Refsgaard JC (Ed.). Kluwer Academic. Ch. 13, pp. 255–278.

Bohlin T, Graebe SF (1995) Issues in nonlinear stochastic grey box indentification. *International Journal of Adaptive Control and Signal Processing* **9**:465–490.

Gneiting T, Balabdaoui F, Raftery AE (2007) Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B* **69**:243–268.

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association* **102**:359–378.

Gupta RS (2008) *Hydrology and Hydraulic Systems*, 3rd Edition. Waveland Press, USA.

Hansen AK, Madsen H, Bauer-Gottwein P, Falk AK, Rosbjerg D (2011) Multiobjective optimization of the management of a waterworks using an integrated well field model. *Hydrology Research*. Accepted.

Harremoës P, Madsen H (1999) Fiction and reality in the modelling world - balance between simplicity and complexity, calibration and identification, verification and falsification. *Water Science and Technology* **39** (9):1–8.

Hendriks-Fransen HJ, KaiserHP, Kuhlmann U, Bauser G, Stauffer F, Müller R, Kinzelbach W (2011) Operational real-time modeling with ensemble kalman filter of variably saturated subsurface flow including stream-aquifer interaction and parameter updating. *Water Resources Research* **47**:W02532.

Jazwinski AH (2007) *Stochastic Processes and Filtering Theory*. Dover Publications, Mineola, New York, USA.

Jonsdottir H, Madsen H, Palsson OP (2006) Parameter estimation in stochastic rainfall-runoff models. *Journal of Hydrology* **326**:379–393.

Kollat J, Reed P (2006) Comparing state-of-the-art evolutionary multi-objective algorithms for long-term groundwater monitoring design. *Advances in Water Resources* **29**:792–807.

Kristensen NR, Madsen H (2003) *Continuous time stochastic modelling - ctsm 2.3 - mathematics guide*. Technical University of Denmark.

Kristensen NR, Madsen H, Jørgensen SB (2004) Parameter estimation in stochastic grey-box models. *Automatica* **40**:225–237.

Madsen H (2008) *Time series analysis*. Chapman & Hall/CRC.

Madsen H, Pinson P, Kariniotakis G, Nielsen HA, Nielsen TS (2005) Standardizing the performance evaluation of short-term wind power prediction models. *Wind Engineering* **29** (6):475–489.

Narasimhan, T. N., Witherspoon, P. A., 1976. An integrated finite difference method for analyzing fluid flow in porous media. *Water Resources Research* **12** (1), 57–64.

Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models part i - a discussion of principles. *Journal of Hydrology* **10** (3):282–290.

Refsgaard JC, van der Sluijs JP, Brown J, van der Keur P (2006) A framework for dealing with uncertainty due to model structure error. *Advances in Water Resources* **29**:1586–1597.

Rosbjerg D, Madsen H (2005) Concepts of hydrologic modeling. In: *Encyclopedia in Hydrological Sciences*. Anderson MG (Ed.). John Wiley and Sons Ltd. Ch. 10.

Rozos E, Koutsoyiannis D (2010) Error analysis of a multi-cell groundwater model. *Journal of Hydrology* **392**:22–30.

Siegfried T, Kinzelbach W (2006) A multiobjective discrete stochastic optimization approach to shared aquifer management: Methodlogy and application. *Water Resources Research* **42**:W02402.

Carrera J (2008) Groundwater: Modeling Using Numerical Methods. In: *Encyclopedia of Water Science*, 2nd Edition. Trimble SW (Ed.). CRC Press. pp. 448–452.

PAPER D

Stochastic simulation and robust optimal management of well fields using Impulse Response Function models

Authors:
G. Dorini, F. Ö. Thordarson, H. Madsen, H. Madsen

Submitted to:
Water Resources Research (2011)

Stochastic simulation and robust optimal management of well fields using Impulse Response Function models

Gianluca Fabio Dorini [1] Fannar Örn Thordarson[1], Henrik Madsen[1]

Abstract

Simulation-based groundwater management models are valuable tools for sustainable exploitation of the natural resource. They allow for pre-evaluations of the impact of management solutions, so that optimality can be assessed. Unfortunately, due to a wealth of uncertainties, simulations often mismatch the actual system, and this ultimately compromises the reliability of the management models. This inherent limit can be partially overcome by using statistical methods to describe the dynamics and to quantify model uncertainty. In this paper we present a multi-period management methodology for a system of pumping wells. The uncertainty in stress-response estimation is handled by employing a special class of Transfer Function-Noise model, known as Predefined Impulse Response Function In Continuous Time. Model parameters are estimated from observed multivariate records using a maximum likelihood method. The method is embedded within a two-steps procedure, whose purpose is to ease the computational burden caused by the dependency between numerical complexity and the number of pumping wells. We consider a chance-constrained optimization problem, which is formalized as convex programming. The optimal solution is computed using Interior Point methods. The methodology is tested on a well field, nearby Copenhagen (DK). The reliability of the management model is proved by testing the accuracy on both estimation set and validation set. An example of head-constrained water supply problem is formulated and solved for different confidence levels of constraints fulfillment. The overall level of uncertainty, i.e. the variance of the objective function is always within the 2% of its own mean value.

Key words:
Simulation, Optimization model, Breakthrough curves, Uncertainty, Time-series, Change-constrained model

[1]Informatics and Mathematical Modelling, Bldg. 305 DTU, DK-2800 Kgs. Lyngby, Denmark

1 Introduction

In many areas of the world, groundwater is the main water resource. Consequences of inadequate management of groundwater aquifers, such as depletion, contamination and land subsidence, are undesirable, particularly for their high social impact (*Tung*, 1987, *Georgakakos and Vlatsa*, 1991). The value of an effective groundwater management methodology is given by the importance of a sustainable use of the natural resource. In literature, groundwater management problems, such as optimal control of aquifer hydraulics, are often addressed using simulation models combined with optimization. Simulation allows for a description of the stress-response relationship in the aquifer (*Tung*, 1986). Optimization utilizes simulation to determine pumping strategies (scheduling), which are best in terms of performance, and feasible in terms of operational constraints fulfillment (*Wagner*, 1999).

Groundwater simulation in management models requires to solve the governing equations, and this is done either by an external simulation model or by using the response matrix approach (see for instance *Das and Datta*, 2001). Simulation models, (also called transient distributed groundwater models, mechanistic models, numerical groundwater models, or distributed parameter groundwater simulation models), are typically deterministic partial differential equations solved numerically by finite difference or finite element methods. Some examples of commercial software packages are MODFLOW (*Harbaugh et al.*, 2000), MIKE-SHE (*Madsen et al.*, 2008) and SUTRA (*Voss*, 1984). The response matrix approach (*Gorelick*, 1983) is based on the concept of Impulse Response Function (IRF) models and linear systems theory. The IRFs describe the response of the water table to an impulse stress (such as pumping, precipitation, etc.) in a set of specified observation points. The response matrix is constructed using IRFs and is computationally more efficient than numeric solvers. However, the use of response matrix is limited to cases where linearity is assessed (e.g. *Tung*, 1986, 1987, *Chang et al.*, 2007). For the vast majority of cases, methodologies deal with complex aquifer systems, employing density dependent transport models, and they are based on the use of transient distributed models (*Wagner and Gorelick*, 1987, *Andricevic*, 1990, *Georgakakos and Vlatsa*, 1991, *Wagner et al.*, 1992, *Morgan et al.*, 1993, *Wagner*, 1999, *McPhee and Yeh*, 2006, *Kalwij and Peralta*, 2006, *Bayer et al.*, 2007, 2010). The advantage of the simulation-optimization approach is that impacts of the management solutions can be evaluated and compared without having to be tested in the real aquifer. Clearly, the reliability of the stress-response estimations is essential for the applicability to real-life case studies.

Uncertainty in groundwater hydraulic management and remediation is mainly due to the difficulty of measuring the spatially-distributed parameters of the

governing equations. Over the past 30 years, lot of research has been carried out to handle the uncertainty limiting the reliability of groundwater management models (e.g. *Das and Datta*, 2001). The earliest methods were commonly based on sensitivity analysis, to study the effects of uncertain parameters variation on the optimal scheduling (*Maddock*, 1974, *Aguado et al.*, 1977, *Willis*, 1979, *Gorelick*, 1982, *Kaunas Jr. and Haimes*, 1985). In the eighties, the mainstream became statistical analysis, with the advantage of capturing the effect of the uncertainty of the parameters on the model sensitivity. With statistical analysis, unknown parameters are assessed using estimation techniques based on field measurements and laboratory experiments (e.g. *Tung*, 1986). Estimates are always encumbered with some uncertainty. Some methodologies mainly focus on hydraulic conductivity or transmissivity (*Tung*, 1987, *Andricevic*, 1990, *Wagner et al.*, 1992, *McPhee and Yeh*, 2006). Spatial variability of conductivity is explicitly considered by *Bayer et al.* (2007) and *Bayer et al.* (2010). In other works, especially those involving remediation problems, uncertainty is handled over the entire parameters set (*Wagner and Gorelick*, 1987). *Georgakakos and Vlatsa* (1991) also considered uncertain boundary conditions. *Chang et al.* (2007) considered uncertain lame coefficient for a problem concerning subsidence control. *Wagner* (1999) and *He et al.* (2008) characterized uncertainty for pollutant concentrations.

In optimal groundwater management, decision variables are commonly pumping rates, which can be time invariant or transient. Performances are estimated using one or more objective functions, which depend on decision variables and hydraulic heads. In order to be taken into account within the decision process, the uncertainty must be propagated from the source to the optimization framework. This can be done in different ways. *Tung* (1986), *Wagner and Gorelick* (1987), *Wagner* (1999), *Chang et al.* (2007) used the first order analysis, the second moment analysis, and the first-order variance-estimation method; *Andricevic* (1990) used the extended Kalman filter; *Georgakakos and Vlatsa* (1991) used the small perturbation method; *McPhee and Yeh* (2006) used the Gaussian quadrature approximation; *Wagner et al.* (1992) used the Monte Carlo sampling; *Morgan et al.* (1993) used the matrix decomposition-based methods; *Kalwij and Peralta* (2006) used the multiple-realization approach; *Bayer et al.* (2007, 2010) used the stack ordering technique. As effect of the uncertainty propagation, both objective function values and constraints fulfillment are random events. In order to make the problem solvable, an equivalent deterministic formulation is required. Probabilistic objective functions are usually transformed into deterministic ones, by taking their expectations. Similarly, constraints can be set to be satisfied with a given probability. This concept is referred to as Chance Constrained optimization (CC), and it has been widely applied in groundwater flow management.

Despite the number and the variety of publications, most of the above listed

methodologies have been successfully tested on synthetic systems, but not validated in real-case study applications. Few exceptions are in water quality (*Kalwij and Peralta*, 2006, *Bayer et al.*, 2010). We argue that this is caused by the inherent difficulty to effectively match real-world systems with transient distributed models. The limitation lies within the spatial discretization of the aquifer domain into cells, and the time discretization, which are necessary for the numerical integration. From this point of view, the response matrix approach would be preferable, as the IRFs are continuous-time models. However, in groundwater management this is hardly exploited, as IRFs are usually determined by performing multiple simulations of unit pumping stress (and contaminant loads for remediation), using again a transient distributed model (*Heidari*, 1982, *Das and Datta*, 2001).

IRFs can alternatively be determined, directly from observed data by means of ARMAX or Transfer Function-Noise (TFN) time series models (see e.g. *Box and Jenkins*, 1970). The structure of TFN models is not necessarily based on the groundflow equation. TFN models estimate an output time series, such as groundwater head in one well, by linearly transforming a multivariate input series, such as pumping stresses in a well field. Uncertainty in output estimation is handled by modeling the residuals of the model as an auto-correlated stochastic process (see for instance *von Asmuth et al.*, 2002). TFN models have been successfully applied in many fields of hydrology, (e.g. *Tankersley et al.*, 1993, *Gehrels et al.*, 1994, *van Geer and Zuur*, 1997). In groundwater simulation, TFN models are often preferred over the use of transient distributed model, not only for their simplicity, but also because their predictions are more accurate (*Hipel and McLeod*, 1994). Furthermore, the noisy component allows for a description of the uncertainty, and also for the extension of the applicability of TFN models to cases where linearity conditions are not fully met.

In this paper we present a TFN model-based groundwater hydraulic management methodology, which is designed to be reliable in real case-study applications. We consider a transient problem of minimum energy use, where a set of linear constraints must be fulfilled (Section 2). The TFN model allows for a CC problem formulation, with quadratic objective function and linear constraints (Section 4). The optimal solution is computed using Interior Point methods (IP), which are extensively employed for practical applications, (see e.g. *Ben-Tal and Nemirovski*, 2001). We use a special class of TFN models, which were developed to deal with hydrologic problems (*von Asmuth et al.*, 2002). In these models, known as Predefined Impulse Response Function In Continuous Time (PIRFICT), the IRFs are defined as simple parametric analytical expressions (Section 3.1). Model parameters are estimated with a maximum likelihood method. Depending on the type of stress, (precipitation, evaporation, pumping wells, rivers fluctuation), different expressions are used (*von Asmuth et al.*, 2008). Here, we propose a particular IRF class of expressions to adapt

PIRFICT models to a field of pumping wells (Section 3.2). The parameter estimation procedure is designed to deal with an arbitrary number of pumping wells (Section 3.3). The methodology is tested using recorded measurements taken at the well field of Søndersø, located northwest of Copenhagen, Denmark (Section 5).

2 The management problem

We consider a well field having N pumping wells. We denote with $q_i(t)$ the pumping stress in well i at time t, and with the vector $\mathbf{q}(t) = (q_1(t), \ldots, q_N(t))^\top$ the stresses in the whole well field. The aquifer response to $\mathbf{q}(t)$ at well i is the piezometric water level $h_i(t)$, and the collection of wells responses is the vector $\mathbf{h}(t) = (h_1(t), \ldots, h_N(t))^\top$. The aquifer's water head response to the pump stresses is described by the groundwater flow equation:

$$S_s \frac{\partial h}{\partial t} = \nabla(\kappa \nabla h) + w + q \tag{1}$$

where S_s and κ are respectively the spatially variable specific storage [L^{-1}] and hydraulic conductivity tensor [LT^{-1}]. The term w accounts for diffuse sources and sinks such as precipitation, evapotranspiration, and river/lakes/sea-level fluctuations. Assuming that both initial condition and time-varying boundary conditions are known, the aquifer's response $\mathbf{h}(t)$ to a given series of stresses $\mathbf{q}(t)$ can be simulated by solving equation (1) in space and time. If the aquifer is confined, the response of well i at time t to a stress j, where $j = 1, \ldots, N$, at time $t' \leq t$ is linear, i.e.

$$\theta_{ij}(t - t') = \frac{\partial h_i(t)}{\partial q_j(t')} \tag{2}$$

where θ_{ij} is the IRF of well i to pumping stress j. Let $t = 0$ be the starting time of the simulation, and $\mathbf{b}(t) = (b_1(t), \ldots, b_N(t))^\top$ be the wells water head for the no-pumping (namely $\mathbf{q}(t) = \mathbf{0}$ for $t \geq 0$); then $h_i(t)$ is given by

$$\begin{aligned} h_i(t) &= \sum_{j=1}^{N} \int_{-\infty}^{t} q_j(\tau)\theta_{ij}(t - \tau)d\tau + b_i(t) \\ &= \sum_{j=1}^{N} \delta_{ij}(t) + b_i(t) \end{aligned} \tag{3}$$

where $\delta_{ij}(t)$ is the drawdown caused by pump rate $q_j(t)$. Simulation allows for the estimation of the impact on the aquifers of different pumping strategies,

and therefore it can be employed as a model for decision support in well field management.

We consider a management model where $\mathbf{q}(t)$ and $\mathbf{h}(t)$ vary along the continuous planning time horizon $t \in [0, T)$. Such horizon is broken down into $K = T/\Delta t$ time intervals of equal duration Δt, denoting the decision time step. In what follows we often refer to as time, both continuous time variable t and discrete time variable k. Decisions are the pump flows over time, and they are taken at the beginning of each time step. More specifically, the decision variable is a discrete-time vector $\mathbf{q}_k = (q_{1k}, \ldots, q_{Nk})^\top$ defining the pump rates during the k-th decision time step, i.e. $q_i(t) = q_{ik}$, for $t \in [(k-1)\Delta t, k\Delta t)$. Similarly, we define $\mathbf{h}_k = (h_{1k}, \ldots, h_{Nk})^\top$, whose i-th component denotes the average head response of well i, along the k-th time step, namely:

$$h_{ik} = \frac{1}{\Delta t} \int_{(k-1)\Delta t}^{k\Delta t} h_i(\tau) d\tau. \tag{4}$$

If the aquifer is confined, the continuous-time IRF $\theta_{ij}(t)$ of equation (2) is turned into a discrete-time function:

$$\theta_{ij,k-k'} = \frac{1}{\Delta t} \int_{(k'-1)\Delta t}^{k\Delta t} \theta_{ij}\big(t' - \tau - (k-1)\Delta t\big) d\tau \leq 0, \tag{5}$$

hence we can estimate the i-th response in discrete-time h_{ik} by replacing the convolution of equation (3) with the expression

$$\begin{aligned} h_{ik} &= \sum_{j=1}^{N} \sum_{k'=1}^{k} q_{k'} \theta_{ij,k-k'+1} + b_{ik} \\ &= \sum_{j=1}^{N} \delta_{ijk} + b_{ik} \end{aligned} \tag{6}$$

where δ_{ijk} is the discrete-time drawdown and the discrete-time no-pumping $\mathbf{b}_k = (b_{k1}, \ldots, b_{kN})$ is obtained by integrating the continuous-time no-pumping $\mathbf{b}(t) = (b_1(t), \ldots, b_N(t))^\top$, using equation (4). Equation (6) can be performed efficiently, as the discrete IRF θ_{ijk} can be determined at once for all i, j, k and then stored into the computer memory. Such approach is known as impact matrix (*Gorelick*, 1983).

The goal of the management is to fulfill a set of constraints, with minimum operational cost. The considered operational cost of pump i at time k is the amount of energy required to lift the water from the aquifer level h_{ik} to the storage level h_s:

$$p_{ik} = \frac{\rho_w \Delta t}{\eta_i} q_{ik} (h_s - h_{ik}) \tag{7}$$

where ρ_w is the density of water [M L^{-3}], and η_i is the efficiency. Assuming linearity (as expressed in equation (6)), the total energy consumption $\mathcal{P} = \sum_{k=1}^{K}\sum_{i=1}^{N} p_{ik}$, is a quadratic function of the stresses $\mathbf{q}_1,\ldots,\mathbf{q}_K$.

Constraints may take the form of linear functions of hydraulic head, stresses and time (e.g. *Ahlfeld and Mulligan*, 2000). Some examples are, stress bounds (e.g. $q_{ik} \leq q_{uk}$, where u is a pump rate constraint); bounds on total stress (e.g. $\sum_{i=1}^{N} q_{ik} \geq D_k$) for water demand ($D_k$) fulfillment; head bound constraints (e.g. $h_{ik} \leq h_{uk}$) for mining and control dewatering or subsidence control; head difference constraints (e.g. $h_{ik} - h_{jk} \geq \Delta_k$) to control salt water or polluted water intrusion within the aquifer. Let N_q and N_h be the number of stress constraints and head constraints, respectively. For $k = 1,\ldots,K$ the constraints are defined as

$$\mathbf{C}\mathbf{h}_k \leq \mathbf{f}_k \quad \text{and} \quad \mathbf{D}\mathbf{q}_k \leq \mathbf{g}_k \tag{8}$$

where $\mathbf{C}$ is a $N_h \times N$ matrix, $\mathbf{f}_k$ are vectors with N_h components, $\mathbf{D}$ are $N_q \times N$ matrices, $\mathbf{g}_k$ are vectors with N_q components. The optimal solution of the management problem is the scheduling, i.e. a sequence of decisions $\mathbf{q}_1,\ldots,\mathbf{q}_K$, minimizing the quadratic function

$$\mathcal{P}^* = \min_{\mathbf{q}_1,\ldots,\mathbf{q}_K} \sum_{k=1}^{K}\sum_{i=1}^{N} \frac{\rho_w \Delta t}{\eta_i} q_{ik}\,(h_s - h_{ik}) \tag{9}$$

subject to linear constraints (6) and (8). Such problem is a quadratic programming problem with linear constraints, whose solution can be computed, for instance, using interior point methods (IP). We remark that in this paper we refer to as pumping wells even in case of simple monitoring wells, i.e. with no pumps installed. Those wells are located in points in which constraints are imposed.

Solving the management problem requires the continuous-time IRFs to be estimated, in order to construct discrete-time IRFs by computing the integral (5). It is important to remark that, following the linearity and time-independent dynamics, such operation only requires to be done once. This is normally done using a transient distributed groundwater model whose accuracy in predictions is limited by several sources of uncertainty. However, if uncertainty cannot be completely eliminated it can be handled using statistical methods, so that management solutions can be assessed taking into account the model accuracy. We do this by using TFN time series models, to determine the IRFs directly from observed data, hence without using transient distributed groundwater models.

3 Modeling uncertainty using Transfer Function Noise time series models

3.1 PIRFICT models

We use TFN models to estimate a groundwater head $h(t)$ in a monitoring well, by linearly transforming N input series of pumping stresses $\mathbf{q}(t)$. Consider a monitoring well in a confined aquifer; as discussed in Section 2, when no pumping, the head follows a certain pattern $b(t)$, which is caused by other sources of stress. In the ideal situation when no source of stress has affected the aquifer for all $t \in (-\infty, +\infty)$, the water level $b(t)$ at that well is a constant b. Let us introduce a source of stress $R(t)$ which is not necessary a pumping stress. This affects the water level in the monitoring well, which deviates from b in time, by an amount given by the convolution

$$h(t) = b + \int_{-\infty}^{t} R(\tau)\theta(t-\tau)d\tau$$

where θ is the IRF to the stress, of the aquifer piezometric level, at the considered well. Calibrating a TFN model consists on determining the shape of the IRF. In time series analysis, IRFs are usually discrete-time rational polynomial expressions (see e.g. *Madsen*, 2008). The PIRFICT models form a special class of TFN, where the IRF is defined as a parametric analytical expression in continuous time. The advantage of the continuous time domain is that both model identification and parameter estimation are independent of the sampling frequency of the observed data. Furthermore, the sampling frequency of the input time series can be irregular. *von Asmuth and Maas* (2001) noticed the similarity between the shape of IRFs and probability density function of continuous random variables. *von Asmuth et al.* (2002) proposed a Pearson type III distribution function for distributed types of stress, such as precipitation, evaporation, and barometric pressure. For other types of stress, such as the influence of pumping wells, or surface water fluctuations, they proposed IRF models that are inspired to physical laws. The IRF of a pumping (or injecting) well, located at distance r from the monitoring well, with pump rate $q(t)$, is inspired to the Hantush formula describing penetrating well in an aquifer of infinite extent, with transmissivity T [L^2T] and storage coefficient S [-], covered by a storage-free aquitard with resistance C [T]:

$$h(t) = b - \int_{-\infty}^{t} \frac{q(t-\tau)}{4\pi T\tau} \exp\left(-\frac{r^2 S}{4T\tau} - \frac{\tau}{CS}\right) d\tau \tag{10}$$

being the convolution between $q(t)$ and a term whose structure is the shape of the proposed IRF:

$$\theta(t) = -\frac{A}{t}\exp\left(-\frac{\beta^2}{\gamma^2 t} - \gamma^2 t\right) \tag{11}$$

where A, γ and β are simply parameters without a physical meaning. Similarly for surface water fluctuations, *von Asmuth et al.* (2008) proposed a parameterized version of the polder function of *Bruggeman* (1999). A similar approach in groundwater simulation can be found in paper by *Tung* (1986), where either the Cooper-Jacob equation, or the Theis equation were proposed as IRF. Also, *Tung* (1987) used the Thiem equation.

The main difference between these approaches and the TFN model approach, lies within the parameter estimation. In fact, for both cases, the IRF parameters were transmissivity and storage coefficient, which were estimated based on pumping tests averaged over a large and representative aquifer volume. Estimation of TFN models is instead exclusively focused in matching the stress-response observation, i.e. ignoring the system's hydrology. Parameters of PIRFICT models are estimated from a time series of observed stresses and water levels at the monitoring well, and the IRF can be used for simulation, i.e.

$$h(t) = b + n(t) + \int_{-\infty}^{t} R(\tau)\theta(t-\tau)d\tau$$

where the term $n(t)$ is the residual [L], namely the deviation between the measurements and the model output, conditioned on the initial value $h(0)$. Ideally are the residuals detected as white noise terms, indicating that the proposed IRF model is the best possible description of the measured output, and the remaining noise is dedicated to fluctuations in the measurements. However, this is seldom the case and the residual series is a combination of the measurement noise and model noise, where the latter can be related to incomplete model structure, undetected input variables or corrupted measurements in the input series. The rigid structure of the IRF models disables the possibility of extending the model, due to the choice of which type of pre-defined curve should be used to approximate the real IRF. Thus, to account for any structural behavior in the residual series, the colored noise $n(t)$ can then be modeled as a stochastic process, given by the stochastic Itô integral

$$n(t) = \int_{-\infty}^{t} \phi(t-\tau)dW(\tau) \tag{12}$$

where $W(t)$ is a continuous white noise (Wiener) process [L] (see e.g. *Oksendal*, 2003). This integral convolution representation was applied by *von Asmuth et al.*

(2002), who proposed an exponential noise model as parametric IRF:

$$\phi(t) = \sqrt{2\alpha\sigma_n^2}e^{-\alpha t} \tag{13}$$

with the parameter α determining the decay rate and σ_n^2 denoting the variance of the residuals. When more than one stress is operating, the overall response of the monitoring well is simply given by the superposition of the responses to individual stresses

$$h(t) = b + n(t) + \sum_i \int_{-\infty}^{t} R_i(\tau)\theta_i(t-\tau)d\tau.$$

Normally, aquifers are affected by all the aforementioned stress sources, which operate in parallel. Consequently, the data utilized for parameter estimation are multivariate time series, of simultaneous measurements of the monitoring well levels and each individual stress. *von Asmuth et al.* (2008) estimated a response model of a well in the Netherlands, to four types of stress: precipitation and evaporation (4 parameters), a pumping well (4 parameters), and fluctuation of a river (4 parameters); for a total of 12 parameters.

3.2 PIRFICT models for well field management

The application of PIRFICT to optimal management of a well field having N pumping wells, requires a estimation of N independent models, i.e. one for each well. Each well model should take into account the pumping stress of all N wells, (including the effect of the pumping well on itself), and all other types of stress. However, since the purpose of optimization is to simulate future scenarios, this requires the future stresses to be know in advance.

Pumping stresses are known in advance, as they are decision variables. Other stresses instead, such as precipitation or temperature, require predictions leading to a consequential increase in uncertainty. Alternatively, we can decide not to explicitly take into account the non-pumping stress inputs, but instead they are embedded in the residual series $n(t)$:

$$h(t) = \sum_{i=1}^{N} \delta_i(t) + b + n(t) \tag{14}$$

where $\delta_i(t)$ is the drawdown caused by pump rate $q_i(t)$, i.e.

$$\delta_i(t) = \int_{-\infty}^{t} q_i(\tau)\theta_i(t-\tau)d\tau. \tag{15}$$

The effect of no-pumping stress, and the effect of all above mentioned sources of uncertainty, can be assessed by performing stochastic simulations of the residual model of equation (12). In particular, assuming that $n(t)$ is known up to time $t = 0$, then each instance of $n(t)$ for $t > 0$ is generated by a random simulation

$$n(t\,|0) = n(0) + \int_0^t \phi(t-\tau)dW(\tau)$$

considering that dW is a Wiener process, and the expression (13), the conditional probability distribution $n(t|0)$ is a normal distribution with mean $n(0)$, and time-dependent variance:

$$\begin{aligned}\mathrm{Var}\{n(t|0)\} =& 2\alpha\sigma_n^2 \int_0^t e^{-\alpha(\tau-t)}d\tau \\ =& \sigma_n^2(1-e^{-\alpha t}).\end{aligned} \tag{16}$$

Consequently, the estimation $h(t\,|0)$ given the stresses $q_1(t),\ldots,q_N(t)$, is normally distributed with mean

$$\begin{aligned}\mathrm{E}\{h(t|0)\} =& \sum_{i=1}^{N}\delta_i(t) + b + n(0) \\ =& \sum_{i=1}^{N}(\delta_i(t) - \delta_i(0)) + h(0)\end{aligned} \tag{17}$$

and variance as expressed in equation (16). We notice that the parameter b is irrelevant. We use this to evaluate time varying intervals entirely containing the stochastic simulation of $h(t\,|0)$

$$h(t|0) \in \left[\sum_{i=1}^{N}(\delta_i(t) - \delta_i(0)) + h(0) \pm \sigma_n\sqrt{1-e^{-\alpha t}}Q_{1-\rho}^{\mathcal{N}(0,1)}\right] \tag{18}$$

at the $1-\rho$ confidence level, where Q denotes the quantile function of standard normal distribution.

For the shape of the IRF $\theta_i(t)$, we propose an alternative expression of the IRF in equation (11). As suggested by the Hantush formula (10), the response to a pumping stress $q_i(t)$ should only depend on the distance between the pumping well and the monitoring well. For all pumping wells, the scaling factor A and the term c, should be the same. Based on that, we derive the following parametric expression:

$$\theta_i(t) = -\frac{A}{t^\beta}\exp\left(-\frac{\beta\lambda_i}{t}\right) \tag{19}$$

where all curves of this family are zero at time $t = 0$, with zero first order derivative; they have a global minimum, the peak delay, which is reached at time $t = \lambda_i$; and finally they asymptotically decay to zero, the same way as $-A/t^\beta \to 0$ for $t \to \infty$; (see Figure 1). The difference between the impulse response of two pumping stresses i, j is specified by the difference in their peak delays λ_i, λ_j. The resulting model has $N + 4$ parameters, i.e.

$$\mathbf{m} = (A, \beta, \lambda_1, \ldots, \lambda_N, \alpha, \sigma_n)^\top.$$

3.3 Parameter estimation

To estimate the parameters in the IRF model the maximum likelihood method is used. This is the same procedure as *von Asmuth et al.* (2002) and *von Asmuth and Bierkens* (2005) applied to calibrate their models, where the aim was to design a methodology to deal with more general cases where the data is irregularly sampled in time. Here, for the sake of brevity, we describe the simpler case of regular sample time interval, and we focus on computational aspects related to the fact that we are dealing with an unspecified number N of pumping wells.

The estimation of the parameters in $\mathbf{m}$ is based on the sequence of the obtained residuals for a given sequence of M instant measurements of the well field activity for the i-th well, $O = [n(t_1), \ldots, n(t_M)]$, with a regular time interval Δt_s. Starting from any initial parameter set $\mathbf{m}_0$, the residuals can be calculated as

$$n(t_i|t_1) = A \sum_{j=1}^{N} \left(\bar{\delta}_j(t_i) - \bar{\delta}_j(t_1) \right) + h(t_1) - h(t_i) \tag{20}$$

where $\bar{\delta}_j(t_i)$ is the j-th drawdown from equation (15) divided by the common scale factor A, namely $\bar{\delta}_j(t_i) = \delta_j(t_i)/A$. Equation (12) with the exponential IRF in equation (13) can be rewritten as (see for instance *von Asmuth and Bierkens*, 2005):

$$\begin{aligned} n(t_i|t_1) =& e^{-\alpha \Delta t_s} n(t_{i-1}|t_1) + \int_{t-\Delta t_s}^{t} \sqrt{2\alpha\sigma_n^2} e^{-\alpha(t-\tau)} dW(\tau) \\ =& e^{-\alpha \Delta t_s} n(t_{i-1}|t_1) + \nu(t_i), \end{aligned} \tag{21}$$

which is an Ornstein-Uhlenbeck process, where $\nu(t_i)$ is the innovation at time instant i. Innovations can be approximately considered to be normally independently distributed random numbers, with mean equal to zero and variance $\sigma_\nu^2 = (1 - e^{-2\alpha\Delta t_s})\sigma_n^2$. From equation (21) a sequence of innovations is given,

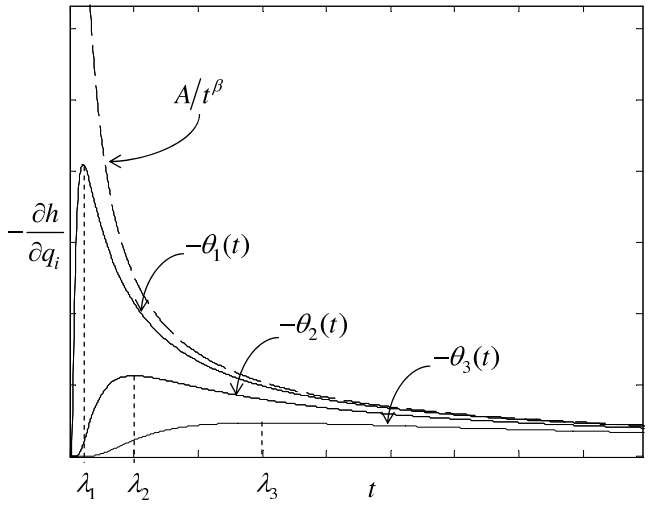

Figure 1: The role of parameters A, β and λ_i in shaping the IRF of equation (19)

$\boldsymbol{\nu}_M = (\nu(t_1),\ldots,\nu(t_M))$, and the conditional likelihood function is obtained as the joint probability density

$$\begin{aligned} L(\mathbf{m};\boldsymbol{\nu}_M) &= p(\nu(t_M)|\boldsymbol{\nu}_{i-1},\mathbf{m}) \\ &= (2\pi\sigma_\nu^2)^{-\frac{M}{2}} \prod_{i=1}^{M} \exp\left(\frac{-\nu^2(t_i)}{2\sigma_\nu^2}\right). \end{aligned} \tag{22}$$

From this, we derive the log-likelihood

$$\begin{aligned} l(\mathbf{m};\boldsymbol{\nu}_M) &= \log\left(L(\mathbf{m};\boldsymbol{\nu}_M)\right) \\ &= -\frac{M}{2}\log\left(2\pi\sigma_\nu^2\right) - \frac{1}{2}\sum_{i=1}^{M}\frac{\nu^2(t_i)}{\sigma_\nu^2}, \end{aligned} \tag{23}$$

which is the objective function to maximize in order to obtain maximum likelihood estimates for the parameters in the IRF models. The standard deviation σ_ν is immediately obtained from the optimality condition $\partial L_\nu / \partial \sigma_\nu = 0$

$$\sigma_\nu^2 = S_\nu^2(\nu(t_i)\,|\mathbf{m},O) = \frac{1}{M}\sum_{i=1}^{M}\nu^2(t_i), \tag{24}$$

hence, the variance of the innovations coincides with its mean square sum $S_\nu^2(\nu(t_i)\,|\mathbf{m},O)$. We notice, by replacing equation (24) into (23), that the maximum of the log-likelihood (23) is attained when σ_n^2 is minimum, therefore parameter estimation is alternatively facilitated by minimizing $S_\nu^2(\nu(t_i)\,|\mathbf{m},O)$.

The model estimation requires $N+3$ independent parameters $A, \beta, \lambda_1, \ldots, \lambda_N, \alpha$ to be estimated.

The model is estimated on a set of observations using the Levenberg-Marquardt optimization algorithm (*Marquardt*, 1963). The algorithm starts from an initial parameter set $\mathbf{m}_0$ and improves on iteratively, until function (24) is minimized. Normally function S_v^2 is non-convex function of $\mathbf{m}$, hence if the initial parameters estimate is distant from the optimum, then any optimization algorithm is likely to get stuck in some local minimum rather than finding the global optimum. Such complexity grows with the number of parameters to estimate, i.e. it grows with the number of pumping stresses N. Consequently, the feasibility of the proposed PIRFICT approach to model well field systems is bounded by the number of wells. This inherent limit can be partially overcome if the Levenberg-Marquardt algorithm is initialized with a good first initial guess for the parameters $\mathbf{m}_0$. This can be done by again solving the problem of maximizing S_v^2, subject to a constraint on the peak delays $\lambda_1, \ldots, \lambda_N$, which are set as function of the distance r_j between the monitoring well and the j-th pumping well

$$\lambda_i = (cr_i)^m \quad i = 1, \ldots, N \tag{25}$$

where c and m are parameters that are common to all pumping stresses. This equals to replace the IRF in equation (19) with the function

$$\widetilde{\theta}_j(t) = -\frac{A}{t^\beta} \exp\left(-\frac{\beta (cr_j)^m}{t}\right). \tag{26}$$

The idea of the peak delays being function of the distance is suggested by the Hantush formula in homogeneous aquifer (10). Clearly, for real case studies, this is only partially realistic, as the aquifers are usually non-homogeneous. On the other hand, the additional constraint in equation (25) reduces the problem complexity by making the degrees of freedom of the variable to optimize independent of N. The simplified problem has in fact only 5 independent parameters A, β, m, c, α. Consequently, using Levenberg-Marquardt algorithm to maximize S_v^2, subject to the constraint in equation (25), is likely to yield a good initial parameter estimates $\mathbf{m}_0$. We call pre-PIRFICT this simplified model. We then improve on $\mathbf{m}_0$ using again Levenberg-Marquardt algorithm to maximize S_v^2, this time without the constraint (25).

The validation of the stochastic model is tested by the verifying the autocorrelation of the innovation series. By this we can visually detect if any model noise remains in the residual series and needs to be accounted for in the model formulation. A model is considered sufficient when it grasps all aspects of the system and the remaining residuals can be considered as a white noise sequence, but white noise terms are independent and Gaussian distributed which indicates that a series of white noise terms should show no autocorrelation.

4 Chance Constrained formulation of the management problem

In this section the PIRFICT methodology is integrated within the management problem described in Section 2, to model the uncertainty in stress-response estimate. We consider an aquifer system having N pumping wells and a multivariate time series of M observations. The residual series for the N models are then described as $\mathbb{O} = (\boldsymbol{n}_1, \ldots, \boldsymbol{n}_N)$. These is the same residuals as obtained in Section 3.3, but here the response is multivariate instead of being univariate. The data in O is used to estimate the parameters $\mathbf{m}_1, \ldots, \mathbf{m}_N$ in all N independent PIRFICT models, according to the procedure described in previous section. The i-th model is identified by notation $\mathbf{m}_i = (A_i, \beta_i, \lambda_{i1}, \ldots, \lambda_{iN}, \alpha_i, \sigma_{n_i})^\top$. Similarily as done in Section 2, we consider a management period T, a decision time steps Δt, and we define a management problem having $K = T/\Delta t$ decision steps. Assuming that $q(t), h(t)$ are known up to time $t = 0$, based on equations (15) and (17) the aquifer heads $\mathbf{b}(t) = (b_1(t), \ldots, b_N(t))^\top$ for the no-pumping, $\mathbf{q}(t) = \mathbf{0}$ for $t > 0$, are given by the convolution

$$b_i(t) = \sum_{j=1}^{N} \int_{-\infty}^{0} q_j(\tau)\theta_{ij}(t-\tau)d\tau - \sum_{j=1}^{N} \int_{-\infty}^{0} q_j(\tau)\theta_{ij}(-\tau)d\tau + h(0). \tag{27}$$

The discrete time heads $\mathbf{b}_1, \ldots, \mathbf{b}_K$ are obtained from $\mathbf{b}(t)$ using formula (4). The discrete time IRFs θ_{ijk} are constructed using equation (5), and the discrete time head-response estimate $\mathbf{h}_1, \ldots, \mathbf{h}_K$ are

$$\begin{aligned} h_{ik} &= \sum_{j=1}^{N} \sum_{k'=1}^{k} q_{k'}\theta_{ij,k-k'+1} + b_{ik} + n_{ik} \\ &= \sum_{j=1}^{N} \delta_{ijk} + b_{ik} + n_{ik} \\ &= \bar{h}_{ik} + b_{ik} + n_{ik} \end{aligned} \tag{28}$$

where $\bar{\mathbf{h}}_k = (\bar{h}_{1k}, \ldots, \bar{h}_{Nk})^\top$ is the expected response for all N wells at time step k. The system formulation is identical to the one in Section 2, except for the additional random element $\mathbf{n}_k$, accounting for all sources of uncertainty. The i-th component n_{ik} is the discretized Ornstein-Uhlenbeck process estimated ac-

cording to equation (21), resulting in the following AR(1) process

$$\begin{aligned} n_{ik} =& \frac{e^{-\alpha_i \Delta t}}{\Delta t} n_{i,k-1} + \frac{1}{\Delta t} \int_0^{\Delta t} \sqrt{2\alpha_i \sigma_{n_i}^2} e^{-\alpha_i \tau} dW(\tau) \\ =& \frac{e^{-\alpha_i \Delta t}}{\Delta t} n_{i,k-1} + \nu_{ik} \\ =& \bar{\alpha}_i n_{i,k-1} + \nu_{ik} \end{aligned}$$

where the discrete-time innovations, ν_{ik}, are normal distributed white noise terms with mean zero and variance

$$\sigma_{\nu_i}^2 = \left(1 - \frac{e^{-2\alpha_i \Delta t}}{\Delta t^2}\right) \sigma_{n_i}^2 = (1 - \bar{\alpha}_i^2) \sigma_{n_i}^2. \tag{29}$$

Thus, the distribution of n_{ki}

$$n_{ik} = \sum_{j=1}^{k} \nu_{ij} \bar{\alpha}_i^{k-j}$$

is normal with mean zero and variance

$$\mathrm{Var}\{n_{ik}\} = (1 - \bar{\alpha}_i^2) \sigma_{n_i}^2 \sum_{i=1}^{k} \bar{\alpha}_i^{2k-2j}. \tag{30}$$

With a stochastic stress-response, the operational cost of pump i at time k is a random variable

$$\begin{aligned} p_{ik} =& \frac{\Delta t \rho_w}{\eta_i} q_{ik} (h_s - \bar{h}_{ik} + n_{ik}) \\ =& \bar{p}_{ik} + \frac{\Delta t \rho_w}{\eta_i} q_{ik} n_{ik} \end{aligned} \tag{31}$$

and, therefore, the total operational cost

$$\begin{aligned} \mathcal{P} =& \sum_{k=1}^{K} \sum_{i=1}^{N} \left(\bar{p}_{ik} + \frac{\Delta t \rho_w}{\eta_i} q_{ik} n_{ik} \right) \\ =& \bar{\mathcal{P}} + \sum_{k=1}^{K} \sum_{i=1}^{N} \frac{\Delta t \rho_w}{\eta_i} q_{ik} n_{ik} \end{aligned} \tag{32}$$

is also a random variable. The stochastic term of equation (32)

$$\begin{aligned}\sum_{i=1}^{N}\sum_{k=1}^{K} q_{ik}n_{ik} &= \sum_{i=1}^{N}\Big(q_{i1}\nu_{i1} + q_{i2}(\bar{\alpha}_i\nu_{i1} + \nu_{i2}) + \cdots \\ &\quad + q_{iK}(\bar{\alpha}_i^{K-1}\nu_{i1} + \cdots + \bar{\alpha}_i\nu_{i,K-1} + \nu_{iK})\Big) \\ &= \sum_{i=1}^{N}\Big(\nu_{i1}\sum_{k=1}^{K} q_{ik}\bar{\alpha}_i^{k-1} + \nu_{i2}\sum_{k=2}^{K} q_{ik}\bar{\alpha}_i^{k-2} \\ &\quad + \nu_{i3}\sum_{k=3}^{K} q_{ik}\bar{\alpha}_i^{k-3} + \cdots + \nu_{iK}q_{iK}\Big)\end{aligned}$$

is a sum of uncorrelated normal distributed numbers with mean zero. Therefore, $\mathcal{P}$ is normal with mean $\bar{\mathcal{P}}$ and variance

$$\mathrm{Var}\{\mathcal{P}\} = \sum_{i=1}^{N}(1-\bar{\alpha}_i^2)\sigma_{n_i}^2\frac{\Delta t^2\rho_w^2}{\eta_i^2}\left(\sum_{k=1}^{K} q_{ik}^2\bar{\alpha}_i^{2k-2} + \sum_{k=2}^{K} q_{ik}^2\bar{\alpha}_i^{2k-4} + \cdots + q_{iK}^2\right) \quad (33)$$

hence the total operational cost in the deterministic case coincides with the expected total operational cost in the stochastic case. Following this, we set as objective of the stochastic optimization problem, the scheduling $\mathbf{q}_1,\ldots,\mathbf{q}_K$ attaining the minimum expected total operational cost

$$\bar{\mathcal{P}}^* = \min_{\mathbf{q}_1,\ldots,\mathbf{q}_K}\sum_{k=1}^{K}\sum_{i=1}^{N}\frac{\Delta t\rho_w}{\eta_i}q_{ik}(h^s - \bar{h}_{ik}) \quad (34)$$

Besides the constraint of eq (28), any feasible scheduling should fulfill the linear constraints of eq (8). Now, stress constraints $\mathbf{D}\mathbf{q}_k \leq \mathbf{g}_k$ can still be imposed as they only affect the decision variables. The head constraints instead

$$\mathbf{C}\mathbf{h}_k \leq \mathbf{f}_k \quad \Rightarrow \quad \mathbf{C}\bar{\mathbf{h}}_k + \mathbf{C}\mathbf{n}_k \leq \mathbf{f}_k \quad (35)$$

are fulfilled probabilistically, $P\{\mathbf{C}\mathbf{h}_k \leq \mathbf{f}_k\} \geq 1-\rho$, as the stochastic component of the j-th constraint $\sum_{i=1}^{N} C_{ji}n_{ik}$, is normal with mean zero and variance

$$\mathrm{Var}\left\{\sum_{i=1}^{N} C_{ji}n_{ik}\right\} = \sum_{i=1}^{N} C_{ji}^2\mathrm{Var}\{n_{ik}\} \quad (36)$$

Hence, as for the total operational cost, the head constraints in the deterministic case, coincides with the expected head constraints in the stochastic case. Although we could impose such expectation constraint $\mathbf{C}\bar{\mathbf{h}}_k \leq \mathbf{f}_k$, it would result in a not robust optimization problem. In fact the optimal scheduling is usually located on some extreme points of the feasible solutions space. Consequently, there are always some linear constraints j that for some times k are satisfied by

the optimal solution with strict equality $\sum_{i=1}^{N} C_{ji}\bar{h}_{ik} = f_{jk}$. In this case the constraints are violated if $\sum_{i=1}^{N} C_{ji}\bar{h}_{ik} > 0$, hence with probability 0.5. Robustness can be achieved by reformulating the problem so that head constraints are set to be fulfilled within given level of confidence. We do that by computing the quantity $I_{jk}(\rho)$ identifying the $1-\rho$ confidence interval of $\sum_{i=1}^{N} C_{ji}n_{ik}$, obtained from equations (35) and (36):

$$\begin{aligned}\sum_{i=1}^{N} C_{ji}n_{ik} \in &\left[\pm\sqrt{\sum_{i=1}^{N} C_{ji}^2 \mathrm{Var}\{n_{ik}\}}\, Q_{1-\rho}^{N(0,1)}\right] \\ &= \left[-I_{jk}(\rho), +I_{jk}(\rho)\right]\end{aligned} \tag{37}$$

We compute vector $\mathbf{I}_k(\rho) = (I_{1k}(\rho), \ldots, I_{N_h,k}(\rho))^\top$ to define the stochastic problem where the optimal scheduling, $\mathbf{q}_1, \ldots, \mathbf{q}_K$ is bounded to fulfill the head constraints for the entire realization of the AR(1) process $\mathbf{n}_1, \ldots, \mathbf{n}_K$, at the $1-\rho$ confidence level. The resulting stochastic formulation of the management problem is the minimization of the expected total operational cost of eq (34), subject to the model constraint of eq (28) and the linear constraints

$$\begin{aligned}\mathbf{C}\bar{\mathbf{h}}_k &\leq \mathbf{f}_k + \mathbf{I}_k(\rho) \\ \mathbf{C}\bar{\mathbf{h}}_k &\leq \mathbf{f}_k - \mathbf{I}_k(\rho) \\ \mathbf{D}\mathbf{q}_k &\leq \mathbf{g}_k.\end{aligned} \tag{38}$$

This formulation is known as chance constrained (CC). Similar CC formulations can be found in the literature; for instance *Tung* (1986) considered an index of confidence of type ρ_{kj}, i.e. depending of time and location.

Even the CC optimization formulation is a quadratic programming problem subject to linear constraints, and therefore the optimal solution $\bar{\mathcal{P}}^*$ can be computed using IP methods. The impact in computational complexity for handling uncertainty is the increase in the number of linear constraints, i.e. from $N_q + N_h$ to $N_q + 2N_h$. It has been shown how uncertainty can be easily handled within the PIRFICT models, and then propagated all the way to the total operational cost $\mathcal{P}^*$, through stochastic optimization. The variance $\mathrm{Var}\{\mathcal{P}^*\}$ of the normally distributed $\mathcal{P}^*$, is calculated using equation (33), and quantifies the level of uncertainty of the optimal well field management solution.

5 Case study

We test the presented methodology in the Søndersø well field, located northwest of Copenhagen (DK), with an annual discharge of 8 mill m^3 of water.

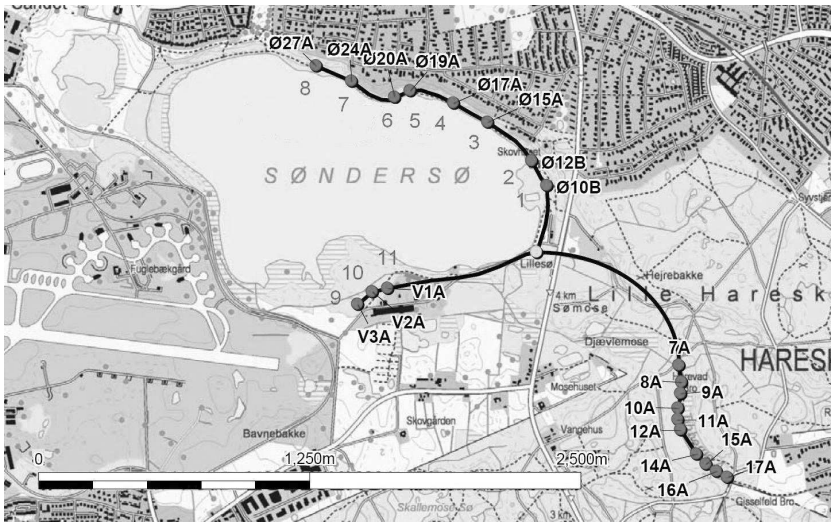

Figure 2: The Søndersø water distribution network.

The system, shown in Figure 2, collects three groups of pumping wells, respectively 9 wells located in the East (labeled starting with 'Ø'), 3 located in the West (V1A,V2A,V3A), and 10 located in the South. The well field covers approximately an area of 4.3×3.7 km. The model contains 8 geological layers (four different clay layers, a sand layer and three different chalk layers). The pumping is mainly done from the chalk layers and partly from the sand layer. Data at disposal are pump rates, at individual wells level for the East and West side, and aggregated for the South part. Hydraulic heads measurements in the wells are also available, except for Ø17A, V3A, and the entire south branch. Based on the data availability, we test a management model for the pump operation in the East and West part. In this exercise we consider the south part as an external well field. The 10 pumps in the south are of the siphon types, capturing water from a superficial layer. Their interaction with the rest of the system, is modest and it can be accounted for in the residuals.

5.1 Estimation results

The available measurements is a dataset of stress-responses sampled every $\Delta t_s = 1$ minute, over a period of approximately 10 months. For a total of 397,625 records. Pump rate measurements are available for all 11 wells of the considered subsystem, whereas hydraulic heads are not available for wells 4 and 9. The 70% of the time series were utilized to calibrate 9 PIRFICT models, i.e. for wells 1, 2, 3, 5, 6, 7, 8, 10, and 11. The remaining 30% of the time series

is used for model validation.

As described in Section 3.3, the parameter estimation procedure consists in two successive optimizations. The first optimization identifies the pre-PIRFICT models; each of them having 5 independent parameters to estimate (A, β, m, c, α). These estimates (Table 1) are then utilized to produce a first estimate $\mathbf{m}_0$ for the second optimization by using formula (25), namely $\lambda_i = (cr_i)^m$ for all $i = 1, \ldots, N$. The second optimization identifies the PIRFICT models; each model having $N + 3$ independent parameters to estimate, $A, \beta, \gamma_1, \ldots, \gamma_N, \alpha$; parameters are listed in Table 2.

Examples of stochastic simulations are in Figure 3, where head responses $h_i(t)$ are represented in terms of confidence interval, and their expectations $\bar{h}_i(t)$ are broken down into individual drawdown levels $\delta_{ij}(t)$, quantifying the contribute of each pump. Note that head responses and drawdown levels in the figure are not in the same scale (i.e. drawdown levels are in a smaller scale).

Table 1: Parameter estimates for the Søndersø well field pre-PIRFICT model, and pump efficiencies

parameter	well1	well2	well3	well5	well6	well7	well8	well10	well11
A [m $\times$ min$^{\beta}$]	0.004	0.003	0.009	0.005	0.0006	0.005	0.003	0.003	0.002
β [-]	1.01	0.958	1.15	1.05	1.08	1.03	0.911	0.967	0.899
m [-]	0.62	0.46	0.62	0.68	1	0.55	0.87	0.45	0.48
c [min$^{1/m}$ $\times$ m^{-1}]	0.0041	0.003	0.0094	0.0047	0.00063	0.0049	0.0026	0.003	0.002
α [-]	0.006	0.002	0.05	0.07	0.01	0.01	0.01	0.004	0.01
σ_n [m]	0.277	0.212	0.54	0.405	0.165	0.378	0.653	0.305	0.346
S_n [m] (calibration)	0.311	0.333	0.544	0.401	0.165	0.385	0.659	0.388	0.349

Table 2: Parameter estimates for the Søndersø well field PIRFICT model.

parameter	well1	well2	well3	well5	well6	well7	well8	well10	well11
A [m $\times$ min$^{\beta}$]	0.004	0.003	0.0001	0.005	0.0007	0.004	0.002	0.003	0.003
β [-]	1.05	1.07	1.56	1.11	1.01	1.24	1.27	1.67	0.914
λ_1 [min]	1e-006	6.73	76.74	62.36	33.15	44.97	28.90	19.01	17.59
λ_2 [min]	20.35	7e-005	42.39	43.44	30.74	42.49	23.17	23.55	17.65
λ_3 [min]	22.98	12.80	4e-007	19.74	18.50	28.88	17.30	22.00	17.73
λ_4 [min]	40.15	17.84	42.25	18.18	11.89	25.41	13.05	20.94	17.82
λ_5 [min]	44.88	19.11	87.25	1e-007	4.59	19.86	11.35	17.19	17.99
λ_6 [min]	29.54	21.73	44.54	9.92	5e-011	17.11	9.55	24.29	25.57
λ_7 [min]	44.61	23.08	127.68	25.51	11.02	9e-006	4.48	15.81	18.49
λ_8 [min]	99.78	67.61	86.82	63.12	19.33	33.76	8e-010	60.20	19.50
λ_9 [min]	30.81	13.46	41.85	24.26	26.00	33.85	15.82	4.79	8.76
λ_{10} [min]	43.75	27.56	138.10	42.02	41.21	34.51	24.86	8e-005	6.26
λ_{11} [min]	57.92	29.15	64.74	55.62	37.24	41.74	23.61	8.62	7e-005
α [-]	0.004	0.002	0.002	0.02	0.009	0.006	0.003	0.008	0.01
σ_n [m]	0.298	0.349	0.496	0.405	0.197	0.385	0.295	0.366	0.345
S_n [m] (calibration)	0.282	0.252	0.45	0.389	0.171	0.385	0.265	0.317	0.297
S_n [m] (validation)	0.251	0.273	0.614	0.301	0.156	0.389	0.27	0.242	0.256
max q [m^3 $\times$ h^{-1}]	81.4	81.4	59.4	81.4	81.4	59.4	54.4	81.4	81.4
η_i [-]	0.7	0.6	0.8	0.7	0.7	0.6	0.7	0.7	0.7

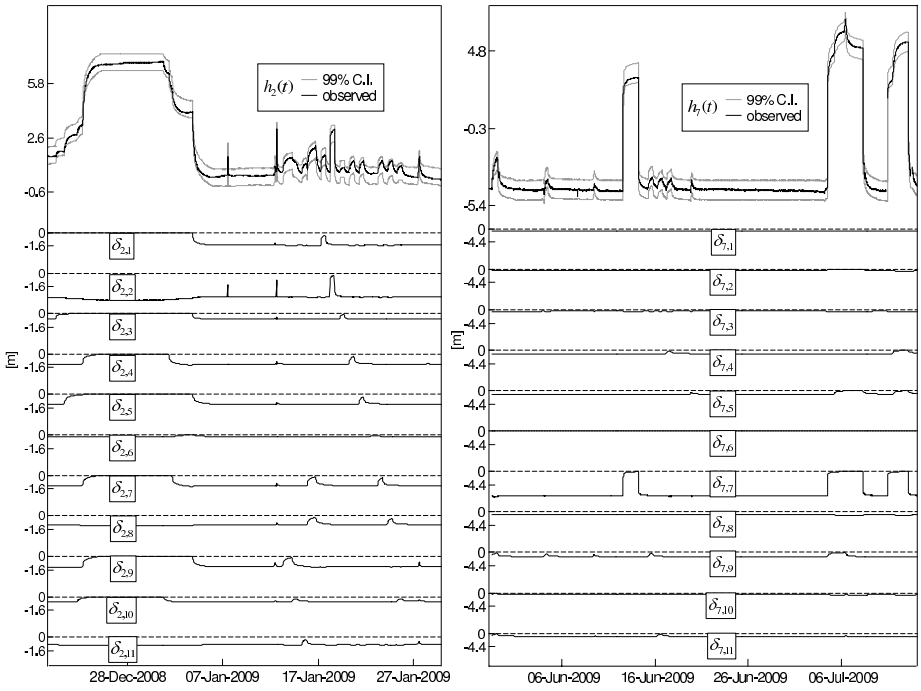

Figure 3: Stochastic simulations of well 2 and 7. The responses $h_2(t), h_7(t)$ are decomposed into drawdown components $\delta_{2,i}(t), \delta_{7,i}(t)$. The dataset utilised for the simuations are part of calibration set for well 2 and part of the validation set for well 7.

The models accuracy in stress-response estimation is assessed using the simulated residuals of the PIRFICT models. The root squared error S_n of the residuals $n_i(t)$ varies within the range of half a meter, whereas the water heads variations $h_i(t)$ can be above 8 meters, as it can be observed in Figure 3. Reliability of the PIRFICT model is also assessed by the fact that S_n is similar in calibration set and validation set (Table 2).

As discussed in Section 3.3, a good first estimate $\mathbf{m}_0$ should be ideally close to the $\mathbf{m}$ attaining the maximum likelihood. A way to assess the goodness of $\mathbf{m}_0$, is therefore to assess the similarity between pre-PIRFICT models and PIRFICT models. We do this by comparing the peaks delay, and the root squared errors. For the PIRFICT models, the peaks delay still tend to grow with the distance, even though they are not constrained to do so (Table 2). In terms of root squared error (see S_n(calibration) on Tables 1 and 2), the pre-PIRFICT models and PIRFICT models perform rather similarly. These similarities suggest that parameters estimate of pre-PIRFICT models are actually a good first

initial parameters estimate for the PIRFICT model.

5.2 Stochastic optimization

We use the estimated PIRFICT models to define and solve a CC management problem for the well field of Søndersø. Since it was not possible to estimate a model for wells 4 and 9, due to the lack of data, the management problem in this exercise is defined on a well field with 7 pumping wells. However, for the sake of consistency with the numbering system adopted throughout this paper, we still consider a well field of $N = 11$ pumping wells, with no pumps and no constraints on wells 4 and 9.

The pump efficiencies η_i and maximum rate, max q_i, are listed in Table 2; the pumps storage level is $h_s = 15$ meters. Note that parameters are missing for wells 4 and 9. The management period covers a time horizon of $K = 30$ decision time steps of a duration of $\Delta t = 3$ hours. Stress constraints $\mathbf{D}\mathbf{q}_k \leq \mathbf{g}_k$ are the pumps range and a water demand of 972 m^3 to be fulfilled in each time step, namely

$$\begin{array}{l} 0 \leq q_{ik} \leq \max q_i \quad i = 1, \ldots, N \\ \sum_{i=1}^{N} q_{ik} \geq 972 \text{ m}^3 \end{array} \tag{39}$$

for all $k = 1, \ldots, K$. Response constraints are lower bounds imposed on wells water level to prevent aquifer from drought withdrawal exceedences, i.e.

$$\begin{array}{ll} h_{ik} \geq -4\text{m} & i = 1, \ldots, 8 \\ h_{jk} \geq 5\text{m} & j = 10, 11 \end{array} \tag{40}$$

for all $k = 1, \ldots, K$. In this simple exercise we assume that the aquifer is undisturbed, i.e. no pumping well was operating before time $t = 0$. The discrete time heads for the no-pumping $\mathbf{b}_k$ are therefore constant. The discrete time IRFs, θ_{ijk}, are constructed using equation (5). The innovations autocorrelation plot in Figure 4, for which the time lag coincides with the decision time step, i.e. 3 hours, shows the validation of the white noise assumption for all 9 PIRFICT models. Traditionally are alternative statistical tools also considered for model checking, where the residual series plays a central role. These methods reveal structural trends in the residual series, which can often be seen by a simple plot of the innovations; where, e.g., any periodicity in the series can be verified with cumulative periodograms, as well as the fit of the model can by quantified by Portmanteau lack-of-fit test (*Madsen*, 2008). However, since the main interest is on the pre-determined decision time step, which is many times greater than the resolution in the available data, the aggregated autocorrelation function for the innovations is considered to give an overview of the lack of fit for the individual decision time step.

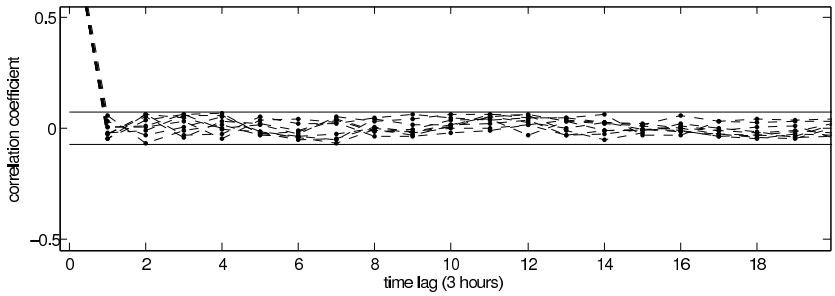

Figure 4: Autocorrelation functions for the innovations of the 9 PIRFICT models of the Søndersø well field system. The solid lines denote the 95% confidence interval, showing that the white noise assumption is valid.

We determine the variances $\sigma^2_{v_i}$ of the innovations ν_{ik} and the coefficients $\bar{\alpha}_i$ using formula (29). Then we determine the variance $\mathrm{Var}\{n_{ik}\}$ of the discrete-time residuals $\mathbf{n}_k$ using formula (30). We then re-formulate the response contraints in equation (40), in terms of robustness with confidence level $1-\rho$ according to equations (36), (37), and (38):

$$\begin{aligned} h_{ik} \pm I_{ik}(\rho) =& h_{ik} \pm \sqrt{\mathrm{Var}\{n_{ik}\}}\, Q^{N(0,1)}_{1-\rho} \geq -4\mathrm{m} \\ h_{jk} \pm I_{jk}(\rho) =& h_{jk} \pm \sqrt{\mathrm{Var}\left\{n_{jk}\right\}}\, Q^{N(0,1)}_{1-\rho} \geq 5\mathrm{m} \end{aligned} \tag{41}$$

for all $k = 1,\ldots,K$. The resulting stochastic formulation of the management problem is the minimization of $\bar{\mathcal{P}}$ (equation (32)), which is a quadratic function of $9 \times 30 = 270$ variables q_{ik}. The problem is subject to $9 \times 30 + 30 = 300$ stress constraints of equation (39), plus $9 \times 30 = 270$ response constraints of equation (40), which double in the robust re-formulation of equation (41). Thus, there are in total $300 + 2 \times 270 = 840$ constraints.

The problem is solved numerically, using the IP methods. Examples of optimal scheduling, with different confidence levels ρ are in Figure 5 and 6. The figures show how the optimal scheduling reduces the stresses whenever the $(1-\rho)$ water head confidence interval intersects the constraint. The reduction in total pump rate is compensated by increasing stress in wells whose head is not close from the constraints. For increasing values of ρ the confidence interval expands in length, reducing the range of feasibility, and ultimately increasing the expected total operational cost $\bar{\mathcal{P}}$; this can be observed in Table 3. Although, as expected, the value of the objective function deteriorates as the confidence level increases, this trade-off is neglectable. In fact, the increment of $\bar{\mathcal{P}}$, between the minimum confidence level $\rho = 1$, coinciding with the deterministic problem formulation and $\rho = .01$ is less than 0.05%. Table 3 also displays the

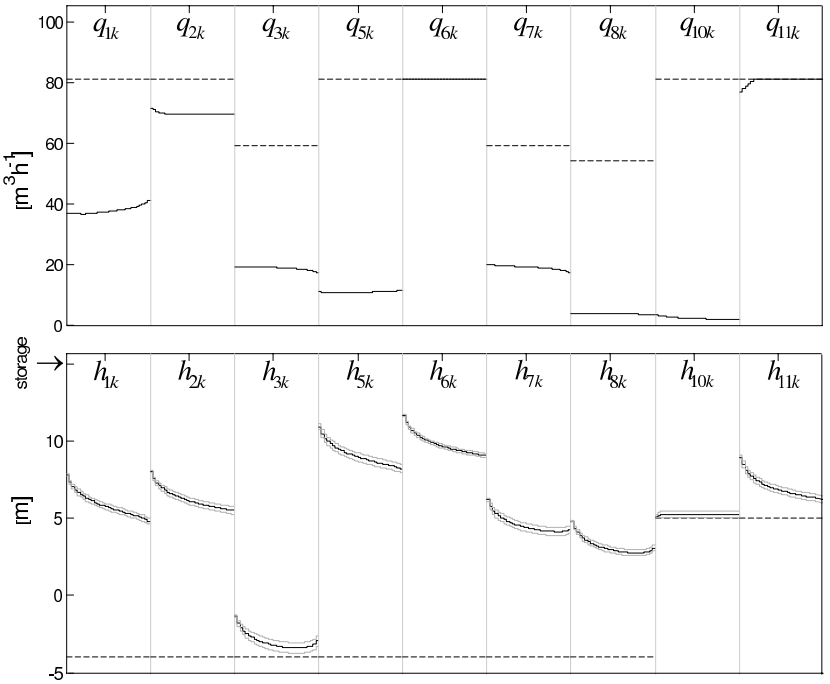

Figure 5: Optimal management of Søndersø well field with low constraint fulfilment confidence level ($1-\rho=0.5$). Top charts: solid lines are the optimal scheduling q_{ik} for all $k=1,\ldots,30$; dashed lines are maximum pumps rates. Bottom charts: solid lines are average responses $\bar{h}_{ik}$ for all $k=1,\ldots,30$, and their $(1-\rho)$ confidence intervals are the grey lines; dashed lines are management constraints.

Table 3: Stochastic optimization of the management problem in the Søndersø well field, for different levels of confidence ρ in head constraints fulfilment.

$1-\rho$	$\bar{\mathcal{P}}$ [Kwh]	$\sqrt{\mathrm{Var}\{\mathcal{P}\}}$ [Kwh]	$\bar{\mathcal{P}}/\sqrt{\mathrm{Var}\{\mathcal{P}\}}$ (%)
0.99	2192.0	41.536	1.895
0.975	2191.7	41.533	1.895
0.95	2191.6	41.531	1.895
0.9	2191.4	41.529	1.895
0.75	2191.2	41.527	1.895
0.5	2191.1	41.527	1.895
0 (deterministic)	2190.9	41.527	1.895

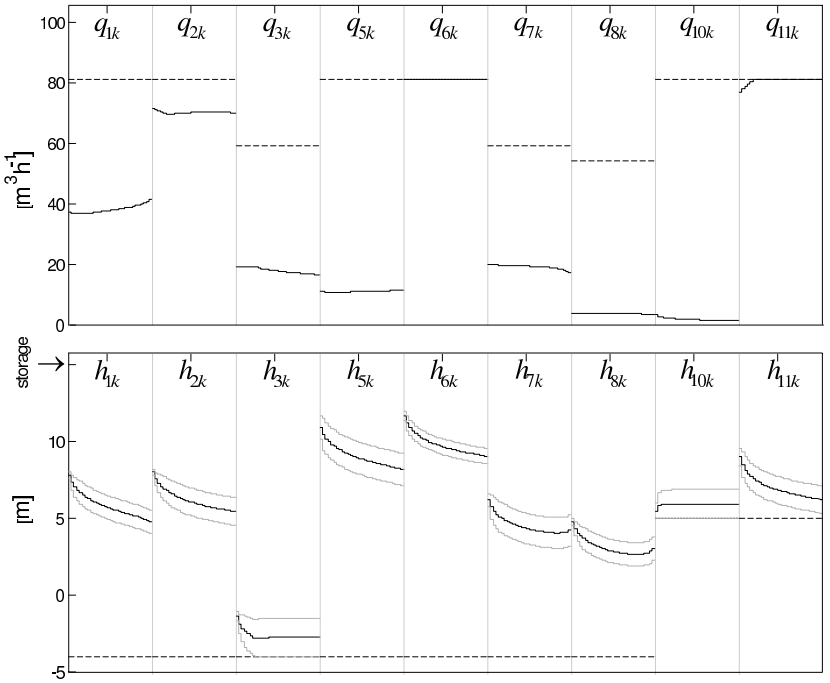

Figure 6: Optimal management of Søndersø well field with high constraint fulfilment confidence level ($1-\rho=0.99$). Top charts: solid lines are the optimal scheduling q_{ik} for all $k=1,\ldots,30$; dashed lines are maximum pumps rates. Bottom charts: solid lines are average responses $\bar{h}_{ik}$ for all $k=1,\ldots,30$, and their $(1-\rho)$ confidence intervals are the grey lines; dashed lines are management constraints.

standard deviation $\sqrt{\mathrm{Var}\{\mathcal{P}\}}$ of the normally distributed total operational cost $\mathcal{P}$. Such measure ultimately quantifies the impact of all sources of uncertainty discussed in Section 2. For this exercise, the uncertainty causes a dispersion of the value of the objective $\mathcal{P}$ which is always around the 2% of its expected value $\bar{\mathcal{P}}$. It can also be seen that, in this case, the trade-off between ρ and $\bar{\mathcal{P}}$ is neglectable, allowing for high confidence level in constraints fulfillment.

6 Discussion and conclusions

This paper describes a groundwater hydraulic management methodology, which is designed for real case-study applications. We consider a problem of minimum expected total energy consumption for a transient scheduling in a system

of pumping wells. The problem is subject to a set of linear constraints, which are functions of hydraulic head, stresses and time. This problem formulation is applicable to a variety of management routines, such as water demand fulfillment, pump rate limits control, mining, dewatering, etc.

Ground water head response to multiple pumping stress is here simulated using the Predefined Impulse Response Function In Continuous Time (PIRFICT). The PIRFICT models are Transfer Function-Noise (TFN) time series models, having the Impulse Response Functions (IRFs) defined as simple parametric analytical expressions which are not related to the systems physics. In this paper we propose a particular IRF class of expressions to adapt PIRFICT models to a system of pumping wells. The methodology is applied to a system of N pumps/observation wells, which requires calibration of N models. Data requirement is limited to wells relative distance and a multivariate dataset of observed hydraulic heads and pump rates. Those types of data can be easily recorded during the routine well pumps operation, hence this methodology does not require specific tests in situ. Although in this work we consider samples taken at regular time intervals, the same methodology can deal with unevenly sampled data records (see *von Asmuth et al.*, 2002, *von Asmuth and Bierkens*, 2005).

Uncertainty in hydraulic heads response to multiple pumping stress is handled within the residual series, modeled as an Ornstein-Uhlenbeck process. The parameter estimation of each of the N PIRFICT models is performed by maximizing a likelihood function of the innovations. For the models proposed, the likelihood function has $N + 3$ independent variables, which is maximized using the Levenberg-Marquardt algorithm. Since the likelihood function is non convex, it has multiple minima, hence the risk for the adopted optimization technique to fail finding the global maximum grows with the number of pumps N. This can be partially overcome if the Levenberg-Marquardt algorithm is initialized with a good first estimate of the parameter set. We provide an initial parameter estimate by firstly reducing the parameterization, and then by applying the same estimation procedure, obtaining a maximum likelihood function with only five independent variables.

Input of the management model are; the planning horizon, the decision time step, the pumps efficiency and capacity, the management constraints. The N estimated PIRFICT models are integrated into impact-matrix-type structure, producing a discrete time stress-response model. The continuous time residuals are integrated into an autoregressive model AR(1). Due to uncertainty in stress-response estimation, both objective functions and head constraint functions values, are random variables with normal distribution. We define an optimization problem, where the objective function is the expected total energy use for pumping, which is quadratic function of the pump rates. Head

constraints are replaced by the extremes of their confidence interval, where the confidence level is set as parameter. Problem of this type are referred to as Chance Constrained optimization (CC). In this form, the CC problem is a quadratic programming problem, and we solve it using Interior Point methods (IP). The IP methods are extensively employed for practical applications, as they are often capable of solving problems within a number of operations not more than polynomial of the problem dimensions. The overall level of uncertainty is quantified by the variance of the objective function.

The methodology is tested using recorded measurements taken at the well field of Søndersø, located northwest of Copenhagen (DK). The management model is defined for a system with $N = 7$ pumping wells. The accuracy of the PIRFICT models is assessed using the simulated residuals of the PIRFICT models. The root squared error of the residuals varies within the range of half a meter, whereas the water heads variations can be above 8 meters. This result is obtained on both calibration set and validation set. Validation of the normal assumptions of the head-response estimation was assessed by considering the autocorrelation of the innovations, evaluated with lag equal to the decision time step.

The management period covers a time horizon of 30 decision time steps, each of a duration of 3 hours. In each time step a water demand must be fulfilled, whilst the water head in the wells are constrained above a minimum level. The resulting CC formulation of the management problem is the minimization of a quadratic function of 270 variables, subject to 840 linear constraints. The problem is solved for different confidence levels of constraints fulfillment, using the IP methods. Although, as expected, the value of the objective function deteriorates as the confidence level increases, this trade-off only causes the objective function to range within a modest range, i.e. about its 0.05%. The overall level of uncertainty, i.e. the variance of the normally-distributed objective function is within the 2% of its own mean value.

Acknowledgements

This work was funded by the Danish Strategic Research Council, Sustainable Energy and Environment Programme, as part of the Well Field optimization project (http://wellfield.dhigroup.com/)

References

Aguado E, Siter N, and Remson I (1977) Sensitivity analysis in aquifer studies. *Water Resour. Res.* **13**:733 – 737.

Ahlfeld DP, and Mulligan AE (2000) *Optimal Management of Flow in Groundwater Systems*. Academic Press, San Diego.

Andricevic R (1990) A real-time approach to management and monitoring of groundwater hydraulics, *Water Resour. Res.* **26**(11):2747 – 2755.

Bayer P, Bürger CM, and Finkel M (2007) Computationally efficient stochastic optimization using multiple realizations, *Advances in Water Resources* **31**(2):399 – 417. doi: 10.1016/j.advwatres.2007.09.004.

Bayer P, de Paly M, and Bürger CM (2010) Optimization of high reliability based hydrological design problems by robust automatic sampling of critical model realizations, *Water Resour. Res.* **46**, W05504, doi: 10.1029/2009WR008081.

Ben-Tal A, and Nemirovski A (2001) *Lectures on Modern Convex Optimization. Analysis, Algorithms, and Engineering Applications*. SIAM, Philadelphia.

Box GEP, and Jenkins MG (1970) *Time Series Analysis: Forecasting and Control*. San Fransisco, California: Holden-Day.

Bruggeman GA (1999) *Analytical Solutions of Geohydrological Problems*.

Chang YL, Tsai TL, and Yang JC (2007) Stochastically Optimal Groundwater Management Considering Land Subsidence. *Journal of Water Resources Planning and Management* **133**(6):486 – 498.

Das A, and Datta B (2001) Application of optimization techniques in groundwater quantity and quality management. *Sadhana* **26**(4):293 – 316.

van Geer FC, and Zuur FA (1997) An extension of Box-Jenkins transfer/noise models for spatial interpolation of groundwater head series. *Journal of Hydrology* **192**:65 – 80.

Gehrels JC, Van Geer FC, and De Vries JJ (1994) Decomposition of groundwater level fluctuations using transfer modelling in an area with shallow to deep unsaturated zones. *Journal of Hydrology* **157**:105 – 138.

Georgakakos AP, and Vlatsa DA (1991) Stochastic Control of Groundwater Systems, *Water Resour. Res.* **27**(8):2077 – 2090.

Gorelick SM (1982) A model for managing sources of groundwater pollution, *Water Resour. Res.* **18**(4):773 – 781.

Gorelick SM (1983) Review of distributed parameter groundwater management modeling methods, *Water Resour. Res. 19*(2):305 – 319.

Harbaugh AW, Banta ER, Hill MC, and McDonald MG (2000) *Modflow-2000, the u.s. geological survey modular groundwater model. User guide to modularization concepts and the groundwater flow process*. Technical report, Open-File Rep. 00-92, U.S. Geological Survey, Washington, D.C.

He L, Huang GH, and Lu HW (2008) A simulation-based fuzzy chance-constrained programming model for optimal groundwater remediation under uncertainty, *Advances in Water Resources* **31**(12):1622 – 1635.

Heidari M (1982) Application of Linear System Theory and Linear Programming to Groundwater Management in Kansas, *Water Resources Bulletin* **18**(6):1003 – 1012.

Hipel KW, and McLeod IA (1994) *Time Series Modelling of Water Resources and Environmental Systems*, Elsevier Sci., New York.

Kalwij M, and Peralta RC (2006) Simulation/optimization modeling for robust pumping strategy design, *Ground water* **44**(4):574 – 582.

Kaunas Jr H, and Haimes YY (1985) Risk management of groundwater contamination in a multiobjective framework, *Water Resour. Res., 21*(11), 1721 – 1730.

Maddock T (1974) The operation of stream-aquifer system under stochastic demands, *Water Resour. Res.* **10**(1):1 – 10.

Madsen H, Gudbjerg J, and Falk AK (2008) A combined groundwater and pipe network model for well-field management, In *MODFLOW and More: Ground Water and Public Policy, Golden, Colorado, USA*, May 19-21.

Madsen H (2008) *Time Series Analysis*, Chapman & Hall/CRC Texts in Statistical Science.

Marquardt DW (1963) An algorithm for least-squares estimation of nonlinear parameters, *J. Soc. Ind. Appl. Math.* **11**(2):431 – 441.

McPhee J, and Yeh WWG (2006) Experimental design for groundwater modeling and management, *Water Resour. Res.* **42**:W02408, doi: 10.1029/2005WR003997.

Morgan DR, Eheart JW, and Valocchi AJ (1993) Aquifer remediation design under uncertainty using a new chance constrained programming technique. *Water Resour. Res.* **29**, 551 – 561.

Øksendal B (2003) *Stochastic differential equations - an introduction with applications*, Springer.

Tankersley CD, Graham WD, and Hatfield K (1993) Comparison of univariate and transfer function models of groundwater fluctuations. *Water Resour. Res.* **29**(10):3517 – 3533.

Tung YK (1986) Groundwater management by chance-constrained model, *Journal of Water Resources Planning and Management* **112**(1):1 – 19.

Tung YK (1987) Multi-objective stochastic groundwater management of nonuniform, homogeneous aquifers, *Water Resources Management* **1**(4):241–254.

von Asmuth JR, and Bierkens MFP (2005) Modeling irregularly spaced residual series as a continuous stochastic process. *Water Resour. Res.* **41**(12):1 – 11, doi: 10.1029/2004WR003726.

von Asmuth JR, and Maas K (2001) The method of impulse response moments: A new method integrating time series-, groundwater- and eco-hydrological modelling, *IAHS-AISH Publ. - Series of Proceedings and Reports* **269**:55 – 58.

von Asmuth JR, Bierkens MFP, and Maas K (2002) Transfer function-noise modeling in continuous time using predefined impulse response functions, *Water Resour. Res.* **38**(12):1287, doi: 10.1029/2001WR001136.

von Asmuth JR, Maas K, Bakker M, and Petersen J (2008) Modeling Time Series of Ground Water Head Fluctuations Subjected to Multiple Stresses, *Ground Water* **46**(1):30 – 34. doi: 10.1111/j.1745-6584.2007.00382.x.

Voss C (1984) *SUTRA: A finite element simulation model for saturated-unsaturated fluid density dependent groundwater flow with energy transport or chemically reactive single species solute transport.*, U.S. Geological Survey Water Resour. Invest. Rep. 84-4369, U.S. Geological Survey, Reston, Va., USA.

Wagner BJ (1999) Evaluating data worth for ground-water under uncertainty management, *Journal of Water Resources Planning and Management* **125**(5):281 – 288, doi: 10.1061/(ASCE)0733-9496(1999)125:5(281).

Wagner BJ, and Gorelick SM (1987) Optimal Groundwater Quality Management Under Parameter Uncertainty, *Water Resour. Res.* **23**(7):1162 – 1174.

Wagner JM, Shamir U, and Nemati HR (1992) Groundwater Quality Management Under Uncertainty: Stochastic Programming Approaches and the Value of Information, *Water Resour. Res.* **28**(5):1233 – 1246.

Willis R (1979) A planning model for the management of groundwater quality. *Water Resour. Res.* **15**(6):1305 – 1312.

PAPER E

Grey box modelling of flow in sewer system with state dependent diffusion

Authors:
A. Breinholt, F. Ö. Thordarson, J. K. Møller, P. S. Mikkelsen,
M. Grum, H. Madsen

Published:
Environmetrics (2011) **22**(8): 946-961

Grey box modelling of flow in sewer systems with state dependent diffusion

Anders Breinholt[1], Fannar Örn Thordarson[2], Jan Kloppenborg Møller[2], Peter Steen Mikkelsen[1], Morten Grum[3], Henrik Madsen[2]

Abstract

Generating flow forecasts with uncertainty limits from rain gauge inputs in sewer systems require simple models with identifiable parameters that can adequately describe the stochastic phenomena of the system. In this paper a simple grey box model is proposed that is attractive for both forecasting and control purposes. The grey box model is based on stochastic differential equations and a key feature is the separation of the total noise into process and measurement noise. The grey box approach is properly introduced and hypothesis regarding the noise terms are formulated. Three different hypotheses for the diffusion term are investigated and compared: one that assumes additive diffusion; one that assumes state proportional diffusion; and one that assumes state exponentiated diffusion. To implement the state dependent diffusion terms Itô's formula and the Lamperti transform are applied. It is shown that an additive diffusion noise term description leads to a violation of the physical constraints of the system, whereas a state dependent diffusion noise avoids this problem and should be favoured. It is also shown that a logarithmic transformation of the flow measurements secures positive lower flow prediction limits, since the observation noise is proportionally scaled with the modelled output. Finally it is concluded that a state proportional diffusion term best and adequately describes the one step flow prediction uncertainty and a proper description of the system noise is important for ascertaining the physical parameters in question.

Key words:
stochastic differential equations, Lamperti transform, parameter estimation, rainfall-runoff, urban drainage

[1]Department of Environmental Engineering, Bldg. 113 DTU, DK-2800 kgs. Lyngby, Denmark
[2]Informatics and Mathematical Modelling, Bldg. 305 DTU, DK-2800 Kgs. Lyngby, Denmark
[3]Krüger, Veolia Water Solutions and Technologies, Gladsaxevej 363, DK-2860 Søborg, Denmark

1 Introduction

The increasing challenges in the urban drainage sector, caused by climate change, stricter environmental regulations and growing urbanisation, have triggered a need for online models to be used for warning and control purposes, (see, for example, Krämer et al, 2007, *Ocampo-Martinez and Puig*, 2009, *Puig et al.*, 2009, *Giraldo et al.*, 2010). However, the inherent uncertainties associated with the model predictions are, rarely accounted for, although there seems to be a consensus from several sources regarding uncertainty in modelling, prediction and simulation with urban drainage models (*Lei and Schilling*, 1996, *Willems and Berlamont*, 2002, *Kuczera et al.*, 2006, *Kleidorfer et al.*, 2009, *Freni and Mannina*, 2010, Deletic et al., 2011).Uncertainty is recognised in input data, in the choice of model structure, parameters and measurements for calibration.

In urban rainfall-runoff modelling, input uncertainties refer to the inadequate measurements of the rain input which is a consequence of spatio-temporal variation of the rainfall events, (*Willems*, 2001, *Vaes et al.*, 2005, *Pedersen et al.*, 2010), as well as errors and biases due to mechanical limitations of the rain gauges, (*Barbera et al.*, 2002, *Molini et al.*, 2005, *Shedekar et al.*, 2009). Rainfall is commonly monitored with the nearest available tipping bucket rain gauges (*Willems*, 2001, *Vaes et al.*, 2005, *Pedersen et al.*, 2010) and as yet only rarely with radars.

Model structure and parameter uncertainty essentially refers to the model design and the parameter estimation method, see the discussion in *Harremoës and Madsen* (1999). Design and performance analysis is typically based on distributed commercial deterministic models like MOUSE (Mike Urban)[1], SWMM[2] and InfoWorks[3]. Such models are often termed white box models, since the considered system is formulated using only the available physical knowledge and any stochasticity in relation to time and space is disregarded. In contrast to the white box models, the black box models are built solely on the consideration of the available data in order to derive a relation between observed input and output. This implies that physical knowledge about the system is ignored and both the model structure and the parameterisation are derived and validated by statistical methods, giving the possibility for developing rigorous stochastic dynamical models that can then provide methods for assessing the prediction uncertainty of the model. Black box models usually provide sufficient short-term predictions when compared to the response time of the system; the system changes are slow, the input errors are significant, but the output errors are small (*Gelfan et al.*, 1999). There are several examples of black

[1] www.dhigroup.com
[2] http://www.epa.gov/nrmrl/wswrd/wq/models/swmm/
[3] www.innovyze.com

box models that have been used to predict flows in sewers, see e.g. *Tan et al.* (1991), *Carstensen et al.* (1998), *El-Din and Schmith* (2002), *Jonsdottir et al.* (2007).

Model-based optimal control of sewer systems presents a case where neither the white box nor the black box approach is ideal. On one hand, a white box model is needed, which is sufficiently accurate to be used for several time steps prediction over wide ranges of state space. On the other hand, black box models provide access to well-developed tools for structural uncertainty identification. The corresponding model development procedure is guaranteed to converge if certain conditions of identifiability of parameters and persistency of excitation of inputs are fulfilled (*Kristensen et al.*, 2004a). In this paper, we use stochastic state space models, also termed grey box models, which consist of a set of stochastic differential equations, (SDEs), describing the dynamics of the system in continuous time and a set of discrete time measurement equations. This methodology provides a way of combining the advantages of black and white box models by allowing prior physical knowledge to be incorporated into the model structure, and subsequently apply statistical methods for parameter estimation and model validation. This typically yields models with both fewer and physically meaningful parameters. As opposed to white box models, parameter estimation in grey box models tends to give more consistent results and less bias, because random effects due to process and measurement noise are no longer absorbed into the parameter estimates, but specifically accounted for by the diffusion and measurement noise terms (*Kristensen et al.*, 2004b). Furthermore, simultaneous estimation of the parameters of these terms provides an estimate of the uncertainty of the model, upon which further model development can be based.

In the present paper a formulation and an estimation of a simple continuous-discrete time stochastic flow model for a sewer system are proposed, which explicitly describe how the measurement and model errors enter into the model. Over the past decades the proposed grey box methodology has been applied in diverse disciplines, e.g. pharmacology (*Tornøe et al.*, 2004), chemical engineering (*Kristensen et al.*, 2004b,a), district heating (*Nielsen and Madsen*, 2006), hydrology (*Jonsdottir et al.*, 2001, 2006), for modelling oxygen concentration in streams (*Jacobsen and Madsen*, 1996), and within urban drainage systems to model pollutant mass to wastewater treatment plant (*Bechmann et al.*, 1999, 2000), flow prediction (*Carstensen et al.*, 1998) and estimation of copper loads in stormwater systems (*Lindblom et al.*, 2007). Generally, the focus of previous studies has been on the physically-based part of the SDE model, the so-called drift term. However, in this article the main focus is on developing the stochastic part of the SDE, the so-called diffusion term, since this part of the SDE is significant for a proper uncertainty description of the flow predictions in an urban drainage system.

Following this introduction, the grey box methodology and important transformations of model states and observations are outlined in Section 2. Section 3 then presents a case study of an urban drainage system with flow measurements affected by both diurnal wastewater variation and rainfall runoff and infiltration inflow. Included here is a description of the catchment area, the data and three model proposals that differ with respect to the diffusion term formulation alone. In Section 4, it is investigated which of the three models best describes the flow predictions and it is checked if that model can be statistically validated. Finally conclusions are drawn in Section 5.

2 Grey box modelling

In order to ease the introduction of the grey box methodology we will begin by presenting the conceptual sewer flow model that later on will be confronted with data from a real catchment area. A conceptual representation of the model is depicted in Figure 1 and a nomenclature of the model is found in Table 1.

2.1 State-space formulation of the conceptual sewer flow model

The commercial physically distributed urban drainage models MOUSE (Mike Urban), SWMM and InfoWorks all build on partial differential equations (PDEs) for pipe flow calculation. However, when calculating the flow at a specific point in the sewer system PDEs can often be simplified by substitution with a set of ordinary differential equations (ODEs), and related to the discrete time observations, using a state-space formulation. It is well known that the rainfall-runoff relationship can be modelled with linear reservoirs in series, (*Jacobsen et al.*, 1997, *Mannina et al.*, 2006, *Willems*, 2010). Hence, the proposed lumped conceptual model for the sewer runoff system displayed in Figure 1 consists of linear reservoirs that are based on ODEs. The first reservoir (S_1) represents the first state variable in the model, receiving runoff from the contributing area A caused by the rainfall registered at the two rain gauges P_{316} and P_{321}. The weighting parameter α is defined to account for the fraction of the measured flow that can be attributed to rain gauge P_{316}. Furthermore, we assume that the measured flow from the contributing area A is fully described by the two rain gauges, implying that the contribution from rain gauge P_{321} is equal to $1 - \alpha$.

The second reservoir (S_2), and correspondingly the second state variable in the two-state model, receives outflow from the first reservoir and diverts it to the flow gauge downstream from the catchment. The purpose of the reservoirs in the model is to simulate the time delay from a rainfall event is being registered

Table 1: Nomenclature of the conceptual flow model.

Symbol	Description	Unit
Inputs:		
$P_{316,t}$	Rain gauge input	m/h
$P_{321,t}$	Rain gauge input	m/h
Rainfall-runoff model parameters:		
A	Impermeable fast runoff area	ha
K	Retention time, fast runoff	h
α	Rain gauge weighting coefficient	-
Wastewater flow model parameters:		
a_0	Average waste water flow	m^3/h
s_1,s_2	Sine constants	-
c_1,c_2	Cosine constants	-
Model states:		
$S_{1,t}$	State of first linear reservoir	m^3
$S_{2,t}$	State of second linear reservoir	m^3
Process noise:		
σ_1	Standard deviance, state 1	m^3
σ_2	Standard deviance, state 2	m^3
Model output:		
Y_k	Observed flow at time step k	m^3/h
Observations:		
$\mathcal{Y}_N$	N number of flow observations	m^3/h
Observation noise:		
ϵ_k	$N(0,S)$	m^3/h
Time:		
k	Time step counter	-
t	Continuous time	h

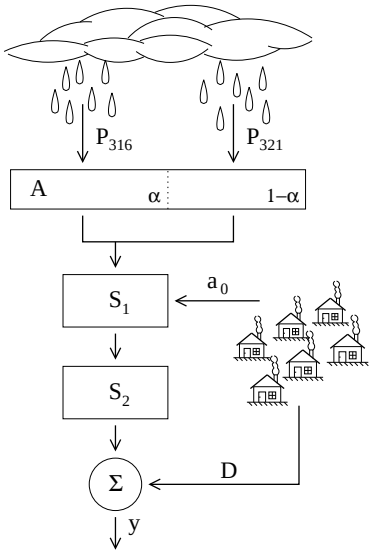

Figure 1: The conceptual model; a system of two linear reservoirs.

at the rain gauges until a corresponding runoff is observed at the location of the flow meter. The time delay is due to both overland runoff time, transportation in the sewer, and in case of heavy rain also internal storage of water in detention basins.

The wastewater flow D is periodic with a diurnal cycle, i.e. in dry weather conditions the observed flow variation is described by the diurnal variation in the wastewater production. The following harmonic function is used:

$$D_k = \sum_{i=1}^{2} \left(s_i \sin \frac{i2\pi k}{L} + c_i \cos \frac{i2\pi k}{L} \right) \tag{1}$$

where L is the period of 24 hours and the parameters s_1, c_1, s_2 and c_2 are non-physical parameters to be estimated.

To fully describe the wastewater flow a constant term for the average dry weather flow a_0 must be added to Equation (1). However, it was decided to attach a_0 to the first state S_1 to secure the physical interpretation of the system, i.e. water is always passing through the system, securing that the reservoirs do not dry out.

By considering the conceptual model displayed in Figure 1 it follows that a

state-space formulation of the model can be described as

$$d\begin{bmatrix} S_{1,t} \\ S_{2,t} \end{bmatrix} = \begin{bmatrix} \alpha AP_{316,t} + (1-\alpha)AP_{321,t} + a_0 - \frac{2}{K}S_{1,t} \\ \frac{2}{K}S_{1,t} - \frac{2}{K}S_{2,t} \end{bmatrix} dt \tag{2}$$

and the observation equation can be formulated as

$$Y_k = \left(\frac{2}{K}S_{2,k} + D_k\right) + \varepsilon_k. \tag{3}$$

The term K in the system Equation (2) represents the mean retention time of the system, i.e. the average time between a rainfall event being registered and the corresponding flow rise being measured by the flow gauge. Diverting the flow through two reservoirs indicates that two retention time coefficients could be used; accordingly, one for the flow from S_1 to S_2 and a second one for the flow from S_2 to the flow measurement station. However, we assume that the two retention times are identical and multiplying with the number of reservoirs in the series, the mean retention time for the flow through the whole sewer system is obtained. It is noted that the second state S_2 appears in the observation equation whereas the first state S_1 is unobserved, i.e. a hidden state. It is furthermore seen that the error between observed and predicted flow is described by the output error term ε_k that is assumed to be a white noise process with $\varepsilon_k \in N(0,S)$, where $N(0,S)$ is a normal distribution with zero mean and variance S.

2.2 Grey box representation of the conceptual model

The model formulation as described by the equations (2) and (3) does not distinguish observation error from input and model structural error. In the grey box methodology this distinction is made by introducing a *diffusion term* also referred to as a process noise term that specifically accounts for model structural deficiencies and input errors in a lumped way. In equation (4) shown below a constant diffusion term has been introduced.

$$d\begin{bmatrix} S_{1,t} \\ S_{2,t} \end{bmatrix} = \underbrace{\begin{bmatrix} \alpha AP_{316,t} + (1-\alpha)AP_{321,t} + a_0 - \frac{2}{K}S_{1,t} \\ \frac{2}{K}S_{1,t} - \frac{2}{K}S_{2,t} \end{bmatrix} dt}_{\text{drift term}} + \underbrace{\begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}}_{\text{diffusion term}} d\boldsymbol{\omega}_t, \tag{4}$$

and the observation equation then changes to

$$Y_k = \left(\frac{2}{K}S_{2,k} + D_k\right) + e_k. \tag{5}$$

The diffusion term adds two standard deviations (σ_1 and σ_2) that account for prediction uncertainty on S_1 and S_2. $\boldsymbol{\omega}_t$ is in this case a 2-dimensional standard

Wiener process, i.e. $d\boldsymbol{\omega}_t \sim \sqrt{dt}N(0,1)$, where $N(0,1)$ is a normal distribution with zero mean and unit variance. The deterministic part of the state equations are referred to as the *drift term*. In Equation (4) the input uncertainty is primarily related to σ_1 because the rain input enters this first reservoir, whereas the model structural uncertainty will appear in both σ_1 and σ_2. The only change in the observation equation (Equation 5) is that ε_k is substituted with e_k because now the total output error (ε_k in Equation (3)) has been divided into a process noise represented by σ_1 and σ_2 and an *observation noise* term (e_k).

In the grey box terminology it is also possible to let the uncertainty on the state predictions depend on the current state level, the inputs or some parameters instead of using a constant diffusion term. In the case of urban drainage systems it seems reasonable to expect that the uncertainty on the state prediction must somehow be related to the rain input. We will return to this in Section 3.2 and now introduce the grey box methodology in its general form:

$$d\boldsymbol{X}_t = \underbrace{\boldsymbol{f}(\boldsymbol{X}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta})}_{\text{drift term}} dt + \underbrace{\boldsymbol{\sigma}(\boldsymbol{X}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta})}_{\text{diffusion term}} d\boldsymbol{\omega}_t \tag{6}$$

$$\boldsymbol{Y}_k = \boldsymbol{h}(\boldsymbol{X}_k, \boldsymbol{u}_k, t_k, \boldsymbol{\theta}) + \boldsymbol{e}_k, \tag{7}$$

where Equation (6) is the *system equation*, describing the dynamic time evolution ($t \in \mathbb{R}_O$) of the physical state of the system in continuous time and Equation (7) is again the *observation equation* that relates the model output to the obeservations $\boldsymbol{Y}_k \in \mathbb{R}^l$ at discrete sampling instants t_k ($k = 1, \ldots, N$) for N number of measurements. Note that in the system equation $\boldsymbol{f}(\cdot) \in \mathbb{R}^n$ represents the drift term and $\boldsymbol{\sigma}(\cdot) \in \mathbb{R}^{n \times n}$ the diffusion term. Here $\boldsymbol{\omega}_t$ is a n-dimensional standard Wiener process. In the system equation, $\boldsymbol{X}_t \in \mathbb{R}^n$ represents the state variables of the model, the input variables are $\boldsymbol{u}_t \in \mathbb{R}^m$ and the parameters are $\boldsymbol{\theta} \in \mathbb{R}^p$. As seen the diffusion term $\boldsymbol{\sigma}(\cdot)$ can be a function of the states, the inputs, the time or some parameter. In the observation equation the observation error term $\boldsymbol{e}_k$ is assumed to be an l-dimensional white noise process with $\boldsymbol{e}_k \in N(\boldsymbol{0}, \boldsymbol{S}(\boldsymbol{u}_k, t_k, \boldsymbol{\theta}))$. It is seen that the observation noise can be a function of the inputs, the time and parameters.

2.3 Parameter estimation

Given the model structure in Equation (6) and Equation (7), the unknown parameters can be determined by finding the parameters that maximise the likelihood function for a given sequence of measurements (*Kristensen et al.*, 2004b).

For time series models, the likelihood function is based on the product of conditional densities, (*Madsen*, 2008). To express the likelihood as product of conditional densities, the rule $P(A \cap B) = P(A|B)P(B)$ is applied, and with a se-

quence of measurements, denoted as $\mathcal{Y}_N = [\boldsymbol{Y}_N, \ldots, \boldsymbol{Y}_0]$, the likelihood function is the joint probability density:

$$L(\boldsymbol{\theta}; \mathcal{Y}_N) = P(\mathcal{Y}_N|\boldsymbol{\theta}) = \left(\prod_{k=1}^{N} P(\boldsymbol{Y}_k|\mathcal{Y}_{k-1}, \boldsymbol{\theta}) \right) P(\boldsymbol{Y}_0|\boldsymbol{\theta}), \tag{8}$$

which is seen by repeated use of $P(A \cap B) = P(A|B)P(B)$. From (8) it is recognised that the likelihood function consists of a product of one-step ahead conditional densities. The likelihood function can only be evaluated if the initial probability density $P(\boldsymbol{Y}_0|\boldsymbol{\theta})$ is known, and all subsequent conditional probability densities can then be assessed by successively solving Kolmogorov's forward equation and applying Bayes' rule, (*Jazwinski*, 2007). In practice, however, this approach is not computationally feasible and an alternative approach is required. Since the system equations in Equation (6) are driven by a Wiener process, which has Gaussian increments, it seems reasonable to assume that the conditional densities can be approximated by Gaussian densities. For linear systems the conditional probabilities in the likelihood function in Equation (8) are Gaussian, but for nonlinear systems this remains an approximation.

The Gaussian density is completely characterised by its mean and covariance of the one step prediction, which are denoted by $\hat{\boldsymbol{Y}}_{k|k-1} = E\{\boldsymbol{Y}_k|\mathcal{Y}_{k-1}, \boldsymbol{\theta}\}$ and $\boldsymbol{R}_{k|k-1} = V\{\boldsymbol{Y}_k|\mathcal{Y}_{k-1}, \boldsymbol{\theta}\}$, respectively, and, by introducing an expression for the innovation formula, $\boldsymbol{\epsilon}_k = \boldsymbol{Y}_k - \hat{\boldsymbol{Y}}_{k|k-1}$ the likelihood function can be rewritten as (*Madsen*, 2008)

$$L(\boldsymbol{\theta}; \mathcal{Y}_N) = \left(\prod_{k=1}^{N} \frac{\exp\left(-\frac{1}{2}\boldsymbol{\epsilon}_k^\top \boldsymbol{R}_{k|k-1}^{-1} \boldsymbol{\epsilon}_k\right)}{\sqrt{\det(\boldsymbol{R}_{k|k-1})}\left(\sqrt{2\pi}\right)^l} \right) P(\boldsymbol{Y}_0|\boldsymbol{\theta}),$$

where the conditional mean and covariance are calculated using a Kalman Filter (KF) for linear models or an Extended Kalman Filter (EKF) for nonlinear models. Finally, the parameter estimates can be obtained by conditioning on the initial values and solving the optimisation problem

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} \{\log(L(\boldsymbol{\theta}; \mathcal{Y}_N|\boldsymbol{Y}_0))\}. \tag{9}$$

In general, it is not possible to optimise the likelihood function analytically, and numerical methods must be applied, (*Kristensen and Madsen*, 2003).

The maximum likelihood method also provides an assessment of the uncertainty for the parameter estimates in Equation (9) since the maximum likelihood estimation is asymptotically normal distributed with mean $\boldsymbol{\theta}$ and covariance matrix

$$\hat{\Sigma}_{\boldsymbol{\theta}} = \boldsymbol{H}^{-1}.$$

The matrix $\boldsymbol{H}$ is the Fisher Information Matrix and is given by

$$h_{ij} = -E\left\{ \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log(L(\boldsymbol{\theta}|\mathcal{Y}_{k-1})) \right\} \qquad i,j = 1,\ldots,p, \tag{10}$$

where in practice an approximation for $\boldsymbol{H}$ is obtained by the observed Hessian h_{ij} evaluated for $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Due to the asymptotic Gaussianity of the estimator in Equation (9), a t-test can be performed to ascertain if the estimated parameters are statistically significant.

When estimating the unknown parameters of the model from a set of data, the continuous-discrete time formulation enables the model to function flexibly with possibilities for varying sample times and missing observations in the data series.

2.4 Transforming the state

To solve the estimation problem, the open source software CTSM[4] (*Kristensen and Madsen*, 2003) is used. Most physical systems have natural constraints in the model structure, e.g. the mass balance in the system cannot be neglected or states need to be positively defined. The restrictions related to positively defined states can partly be dealt with by state dependent diffusion terms in the SDEs. However, this requires a higher order KF which has not been implemented in CTSM, because it was shown to become numerically unstable (*Vestergaard*, 1998). Hence, it is not directly possible to estimate parameters in models with state dependent diffusion terms. To obtain efficiency and numerical stability in the estimation, a transformation of the SDEs is required to generate a new set of equations, where the diffusion term can be independent of the state variable, (*Baadsgaard et al.*, 1997).

The procedure of transforming a general SDE into a form with state independent diffusion term is frequently referred to as the Lamperti Transform, (*Iacus*, 2008). Existence is only guaranteed for one-dimensional diffusion processes, whereas for multi-dimensional diffusion processes, existence depends on the structure of the diffusion term, (*Luschgy and Pagés*, 2006). The one-dimensional diffusion is the simplest case of a state dependent diffusion term in SDEs and only the univariate transformation is considered here. For the multivariate transform, we refer to *Møller and Madsen* (2010).

For any given t assume that the drift term $f_i(\cdot) = f_i(\boldsymbol{X}, \boldsymbol{u}, \boldsymbol{\theta})$, and the diffusion term $\sigma_{ii}(\cdot) = \sigma_{ii}(X_i, \boldsymbol{u}, \boldsymbol{\theta})$, $\sigma_{ij} = 0$ for $i \neq j$, then the SDE for the transformed state

[4] Continuous-Time Stochastic Modelling - www.imm.dtu.dk/ctsm

$Z_i = \phi(X_i) = \phi$ is obtained by Itô's formula (*Øksendal*, 2003):

$$dZ_i = \Big(\frac{\partial \phi}{\partial t} + f_i(\cdot)\frac{\partial \phi}{\partial X_i} + \frac{\sigma_{ii}^2(\cdot)}{2}\frac{\partial^2 \phi}{\partial X_i^2}\Big)dt + \sigma_{ii}(\cdot)\frac{\partial \phi}{\partial X_i}d\omega_i, \tag{11}$$

where ϕ is a twice continuously differentiable function for $(t, X_i) \in (\mathbb{R}_+, \mathbb{R})$. Focusing on the diffusion term in the transformed SDE in Equation (11) shows that the state dependency can be removed from the equation by solving

$$\frac{1}{\sigma_{ii}(\cdot)} = \frac{\partial \phi}{\partial X_i},$$

and the Lamperti transform for the ith state becomes

$$\begin{aligned} Z_i =& \phi(t, X_i) = \int d\phi(t, \xi)\Big|_{\xi = X_i} = \int \frac{\partial \phi}{\partial \xi} d\xi \Big|_{\xi = X_i} \\ =& \int \frac{d\xi}{\sigma_{ii}(\xi, \boldsymbol{u}_t, t, \boldsymbol{\theta})}\Big|_{\xi = X_i}. \end{aligned} \tag{12}$$

The Lamperti transform in Equation (12) provides a system equation with a state independent diffusion term, but the parameters are the same as in the original SDE and the model is still describing the same input-output relationship. Thus, considering a transformation for all system equations in a model, the transformed grey box model is written

$$d\boldsymbol{Z}_t = \widetilde{\boldsymbol{f}}(\boldsymbol{Z}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta})dt + \widetilde{\boldsymbol{\sigma}}(\boldsymbol{u}_t, t, \boldsymbol{\theta})d\boldsymbol{\omega} \tag{13}$$

$$\boldsymbol{Y}_k = \widetilde{\boldsymbol{h}}(\boldsymbol{Z}_k, \boldsymbol{u}_k, t_k, \boldsymbol{\theta}) + \boldsymbol{e}_k, \tag{14}$$

where $\boldsymbol{Z}$ is a vector including the transformed states and the function $\widetilde{\boldsymbol{f}}$ is a description for the drift terms of the transformed state space model. $\widetilde{\boldsymbol{h}}$ represents the new observation equation, but now as a function of the transformed states, and $\widetilde{\boldsymbol{\sigma}}$ is a state independent diffusion term.

2.5 Example of the Lamperti transform

In what follows the properties of the Lamperti transform will be exemplified and later applied in a case study. The notation for the SDE is simplified by omitting input dependencies for the diffusion, since focus is on state dependency. Hence the ith SDE of the system equation in Equation (6) is written as

$$dX_i = f_i(\boldsymbol{X}, \boldsymbol{u}, \boldsymbol{\theta})dt + \sigma_{ii}(X_i, \boldsymbol{\theta})d\omega. \tag{15}$$

The drift term is assumed to be linear. The function f_i can then be separated into two terms, one describing the linear relation to the state (a_i) and a second term (b_i) counting for the relation to any other variable influencing the state X_i, i.e. the input variables $\boldsymbol{u}$ and the remaining states $\boldsymbol{X}^*$, where $\boldsymbol{X}^* = \boldsymbol{X} \setminus X_i$. Using Equation (15), the ith SDE becomes

$$dX_i = \big(b_i(\boldsymbol{X}^*, \boldsymbol{u}, \boldsymbol{\theta}) + a_i(\boldsymbol{\theta})X_i\big)dt + \sigma_{ii}(X_i, \boldsymbol{\theta})d\omega. \tag{16}$$

The focus is now on the diffusion term σ_{ii} while the drift term is considered as displayed in Equation (16). Only the system equation is considered because the observation equation remains unchanged.

Example: $\sigma_{ii}(\cdot) = \sigma_i X_i^{\gamma_i}$

One of the simplest diffusion formulations in SDEs is to assume linear dependency between the state and corresponding noise, but linearity is not always a satisfactory state dependency. Therefore, the diffusion is a function of the state to the power of γ_i, where, for now, γ_i is arbitrary. The system equation then becomes

$$dX_i = \big(b_i(\boldsymbol{X}^*, \boldsymbol{u}, \boldsymbol{\theta}) + a_i(\boldsymbol{\theta})X_i\big)dt + \sigma_i X_i^{\gamma_i} d\omega, \tag{17}$$

where σ_i is a constant term. According to the Lamperti transform in Equation (12), the function $\sigma_{ii}(\cdot) = \sigma_i X_i^{\gamma_i}$ should be considered to obtain the transformed state Z_i, but since σ_i is a constant and not influencing the result of the integration, it can be neglected in the transformation. Consequently, σ_i remains in the system equation in Equation (17) and only the part of the diffusion term with state X_i involved is reflected in the state transformation.

Using Equation (12), the Lamperti transform for the SDE in Equation (17) is then

$$Z_i = \phi(t, x_i) = \int \frac{d\xi}{\xi_i^{\gamma_i}}\bigg|_{\xi = X_i} = \frac{X_i^{1-\gamma_i}}{1-\gamma_i} \Leftrightarrow X_i = \big((1-\gamma_i)Z_i\big)^{\frac{1}{1-\gamma_i}}. \tag{18}$$

To obtain the SDE of the transformed state Itô's formula is applied, as described in Equation (11), but here it utilises both the first and second derivatives of the transformed state Z_i with respect to the original state X_i, as well as the first time derivative of the transformed state. For the transformation in Equation (18), the derivatives in Equation (11) become

$$\frac{\partial \phi}{\partial X_i} = \phi_x = \frac{1}{X_i^{\gamma_i}} \qquad \frac{\partial^2 \phi}{\partial X_i^2} = \phi_{xx} = -\frac{\gamma_i}{X_i^{\gamma_i - 1}} \qquad \frac{\partial \phi}{\partial t} = \phi_t = 0,$$

and Itô's formula then gives the transformed state:

$$\begin{aligned} dZ_i &= \Big(\phi_t + \phi_x f_i + \frac{1}{2}\phi_{xx}\sigma_i^2\Big)dt + \phi_x \sigma_i d\omega \\ &= \Big(0 + \frac{b_i(\cdot) + a_i(\cdot)X_i}{X_i^{\gamma_i}} + \frac{1}{2}\Big(-\frac{\gamma_i}{X_i^{\gamma_i+1}}\Big)\sigma_i^2 X_i^{2\gamma_i}\Big)dt + \frac{\sigma_i X_i^{\gamma_i}}{X_i^{\gamma_i}}d\omega \\ &= \Big(\frac{b_i(\cdot)}{X_i^{\gamma_i}} + a_i(\cdot)X_i^{1-\gamma_i}) - \frac{1}{2}\gamma_i\sigma_i^2 X_i^{\gamma_i-1}\Big)dt + \sigma_i d\omega. \end{aligned} \tag{19}$$

Substitute the state transformation in Equation (18) into the transformed SDE in Equation (19) and obtain,

$$\begin{aligned} dZ_i =& \Big(\frac{b_i(\cdot)}{\big((1-\gamma_i)Z_i\big)^{\frac{\gamma_i}{1-\gamma_i}}} + a_i(\cdot)\big((1-\gamma_i)Z_i\big)^{\frac{1-\gamma_i}{1-\gamma_i}} \\ &- \frac{1}{2}\gamma_i\sigma_i^2\big((1-\gamma_i)Z_i\big)^{\frac{\gamma_i-1}{1-\gamma_i}}\Big)dt + \sigma_i d\omega \\ =& \Big(b_i(\cdot)\big((1-\gamma_i)Z_i\big)^{-\frac{\gamma_i}{1-\gamma_i}} + a_i(\cdot)(1-\gamma_i)Z_i \\ &- \frac{1}{2}\frac{\gamma_i}{1-\gamma_i}\sigma_i^2 z^{-1}\Big)dt + \sigma_i d\omega \\ =& \widetilde{f}_i(\boldsymbol{Z},\boldsymbol{u},\boldsymbol{\theta})dt + \sigma_i d\omega, \end{aligned} \tag{20}$$

corresponding to the ith state in the transformed system equation in Equation (13).

By setting γ_i equal to one, a linear state dependency in X_i can be obtained by applying Equation (12),

$$Z_i = \log(X_i) \Leftrightarrow X_i = e^{Z_i}. \tag{21}$$

The Lamperti transform for a SDE with a diffusion term that is linearly dependent on the state is the logarithmic transform, (or log-transform), since the integral in the Lamperti transform results in a logarithmic relation between the original state and the transformed one. To find the SDE of the transformed state, Equation (11) is again applied to obtain

$$\begin{aligned} dZ_i &= \Big(b_i(\cdot)e^{-Z_i} + a_i(\cdot) - \frac{1}{2}\sigma_i^2\Big)dt + \sigma_i d\omega \\ &= \widetilde{f}_i(\boldsymbol{Z},\boldsymbol{u},\boldsymbol{\theta})dt + \sigma_i d\omega. \end{aligned} \tag{22}$$

Notice that the diffusion parameters σ_i and γ_i in Equation (20) and Equation (22), as well as the model parameters in $b_i(\cdot)$ and $a_i(\cdot)$ are unaffected by the

transformation. Hence, the estimated parameters in the transformed model can be directly implemented into the original model.

To estimate the γ_i parameters, a restriction is required to obtain proper prediction intervals for coverage of the variation in the observations. With $\gamma_i = 0.5$ the state has a positive probability of reaching zero if the input parameters are small compared to the diffusion parameters, (*Iacus*, 2008, *Møller and Madsen*, 2010) and the EKF is not suited for such distributions, whilst for $\gamma_i > 1$ existence and uniqueness of the system are not guaranteed, (*Øksendal*, 2003). Thus, the γ_i parameters need to take values between 0.5 and 1 during estimation.

2.6 Transforming the observations

The implicit assumption of using a constant observation noise term is that the observation noise is independent of states. However, for many physical systems, this is unrealistic and a noise term that increases proportionally with the output is more appropriate, i.e.

$$\boldsymbol{Y}_k = \boldsymbol{h}(\boldsymbol{X}_k, \boldsymbol{u}_k, t_k, \boldsymbol{\theta})\boldsymbol{\epsilon}_k,$$

where ϵ is log-normally distributed and the observation functions $\boldsymbol{h}$ are the same as shown in Equation (7). Consequently, the observation noise is scaled with the size of the measured model output. This is beneficial because studies of flow meter uncertainty have shown that measurement uncertainty increases proportional with the flow magnitude, (*Bertrand-Krajewski et al.*, 2003).

One of the benefits of expressing the observation equation with an additive Gaussian noise, as in Equation (7), is that the assumption of Gaussianity for the residuals enables the use of the EKF and statistical tests to verify the proposed model, (more regarding model validation in the following section). CTSM utilises these tests and in the implementation only additive noise terms in the observation equations are allowed. Thus, to separate the noise term from the model, where the noise is multiplicative and log-normal distributed, a logarithmic transform of the measurements is required:

$$\begin{aligned}\log(\boldsymbol{Y}_k) &= \log\left(\boldsymbol{h}(\boldsymbol{X}_k, \boldsymbol{u}_k, t_k, \boldsymbol{\theta})\boldsymbol{\epsilon}_k\right) \\ &= \log\left(\boldsymbol{h}(\boldsymbol{X}_k, \boldsymbol{u}_k, t_k, \boldsymbol{\theta})\right) + \log\left(\boldsymbol{\epsilon}_k\right) \\ &= \log\left(\boldsymbol{h}(\boldsymbol{X}_k, \boldsymbol{u}_k, t_k, \boldsymbol{\theta})\right) + \boldsymbol{e}_k. \end{aligned} \tag{23}$$

The log-transformed observations can then be applied in CTSM.

2.7 Model validation

One of the main aspects of the grey box modelling framework is its predictive ability, which implies that the output errors are examined for any systematic pattern for further extension of the model. Several statistical tools are utilised for the validation procedure, which all have their own properties for identifying the lack of fit in the model. The statistical tools used in the paper are all well described in *Madsen* (2008).

The model residuals are useful for the validation. The general assumption for the residuals for an adequate model is that they are white noise. Plotting the sample autocorrelation function (ACF), and the sample partial autocorrelation function (PACF) for the residuals will show if the residuals eventually are autocorrelated. In the frequency domain, the cumulative periodogram is useful for detecting the deviation from the white noise assumption for the residuals. With the cumulative periodogram, any hidden periodicities, including seasonality, in the residuals can be detected. For more details on the cumulative periodogram, see *Madsen* (2008) and *Priestley* (1981).

3 Case study and model proposals

The grey box methodology is applied to find a satisfactory flow model for a sewer system. As already seen in Section 2.2 the proposed model has a rather limited physical structure, and therefore the advantages of adequately formulating the diffusion term of the SDEs to cope with model deficiencies and input uncertainties will be emphasised.

3.1 Catchment, drainage system and data

Figure 2 gives an overview of the study catchment, which is situated in the north-western part of greater Copenhagen in Ballerup Municipality. The total area is 1,320 ha. Most of the catchment area (93%) utilises a separate system with two parallel pipes for wastewater and stormwater, while the remaining 7% is served by a combined sewer system in which both wastewater and stormwater flow through the same pipe. A significant amount of infiltration inflow into the sewer network is taking place, probably due to worn-out pipes and faulty connections. A flow meter has been installed downstream of the catchment area to attempt to ascertain the extent of this leakage. The flow meter is a semi-mobile ultrasonic Doppler type. It is placed in an interception pipe with a diameter of 1.4 m. The flow meter logs every 5 minutes.

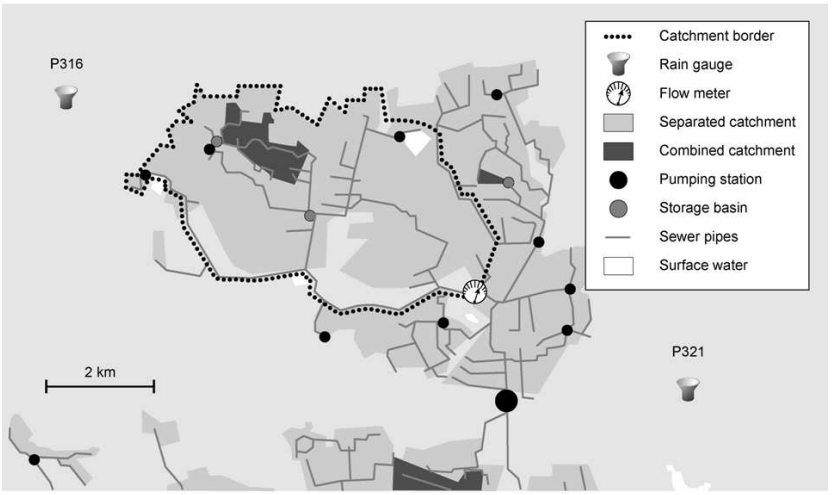

Figure 2: The Ballerup catchment area.

There are around 50,000 inhabitants living inside the catchment area, which is one of several sub-catchment areas that diverts water to the second largest Wastewater Treatment Plant (WWTP) in Denmark, called Avedøre WWTP. There are a couple of small pumping stations and one storage basin inside the catchment area, with an approximate capacity of 4000 m^3. The two closest rain gauges from the national Danish tipping bucket network, (*Jørgensen et al.*, 1998), indicated $P316$ and $P321$ in Figure 2, have a 0.2 mm resolution and are located outside the studied catchment area, approximately 12 km apart.

A nearly three month period, (April 1st - June 21st, 2007) is used for estimation. The measured precipitation varies considerably from one rain gauge to the other and spatio-temporal rainfall variation is clearly identified. This is illustrated in Figure 3 that shows the accumulated precipitation measured at each rain gauge, (P_{316}) and (P_{321}) plotted on a log scale. If a given rainfall registration at the two gauges is separated by more than one hour, they are considered to be separated events. Note how this distinction results in some rainfall events being registered at only one of the rain gauges.

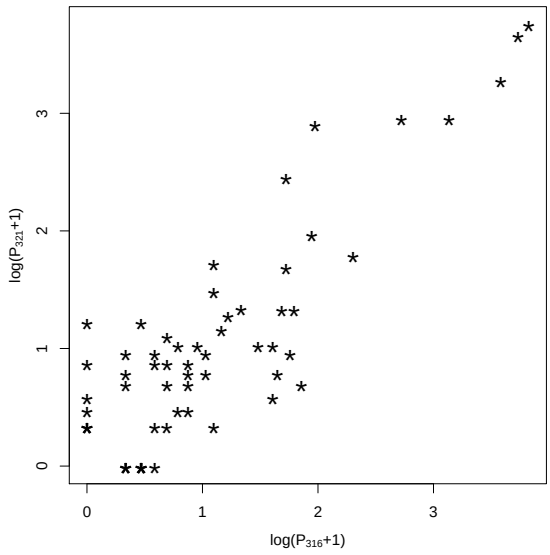

Figure 3: Correlation between the two rain gauges. The measured precipitation varies considerably from one rain gauge to the other.

3.2 Diffusion term proposals

Comparing the drift term of the SDE in Equation (16) with the drift term of the system equation in Equation (4), it is seen that the flow model can be rewritten

$$d\begin{bmatrix} S_{1,t} \\ S_{2,t} \end{bmatrix} = \begin{bmatrix} b_1(\boldsymbol{u}_t, \boldsymbol{\theta}) + a_1(\boldsymbol{\theta}) S_{1,t} \\ b_2(S_{1,t}, \boldsymbol{\theta}) + a_2(\boldsymbol{\theta}) S_{2,t} \end{bmatrix} dt + \boldsymbol{\sigma}(\boldsymbol{S}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta}) d\boldsymbol{\omega}_t \,, \tag{24}$$

where

$$a_i(\boldsymbol{\theta}) = a_i(K) = -\frac{2}{K} \qquad \text{for } i = 1,2$$

$$b_1(\boldsymbol{u}_t, \boldsymbol{\theta}) = b_1(P_{316,t}, P_{321,t}, A, \alpha, a_0) = \alpha A P_{316,t} + (1-\alpha) A P_{321,t} + a_0$$

$$b_2(S_{1,t}, \boldsymbol{\theta}) = b_2(S_{1,t}, K) = \frac{2}{K} S_{1,t}.$$

The observation equation remains the same for all model proposals and we refer to the grey box model represented by Equation (4) and Equation (5) where in the following only the diffusion matrix $\boldsymbol{\sigma}(\boldsymbol{S}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta})$ is modified in order to obtain an improved description of the flow uncertainty. The models are estimated on a 15 minutes time resolution.

Model 1 The first model proposal is a model where the diffusion term is considered constant, corresponding to the model presented in Section 2.2. Model 1 is then represented with the diffusion matrix

$$\boldsymbol{\sigma}(\boldsymbol{S}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta}) = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix},$$

and the diffusion parameters (σ_1, σ_2) are estimated as described in Section 2.3. Since the diffusion in the model is state independent, no transformation of the states is required to estimate the model parameters.

Model 2 The drift term of the model is driven by transient rain events, implying that most of the time the flow in the sewer system consists of wastewater flow only. In that case, the variance of the diffusion term is expected to be rather small, but when a rain event occurs the variance is expected to increase significantly, due to the uncertainty in the actual rain input to the system. It is furthermore anticipated that the uncertainty increases with the magnitude of the rainfall, (both duration and magnitude), which is captured by state dependent diffusion.

Introducing a state dependent diffusion term has the desired implication that the diffusion is scaled with the state magnitude. This makes physical sense since the diffusion terms, (especially the first one), primarily represent the uncertainty in the rain input, and therefore should not contribute any uncertainty to the output, (the flow), in dry weather periods. Another implication is that the risk of receiving negative state values is avoided as discussed in Section 2.5. Model 2 is represented with the state proportional diffusion matrix

$$\boldsymbol{\sigma}(\boldsymbol{S}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta}) = \begin{bmatrix} \sigma_1 S_{1,t} & 0 \\ 0 & \sigma_2 S_{2,t} \end{bmatrix}.$$

With the addition of state dependency, it is expected that the diffusion parameters will be reduced, since the state variation is adjusted with the state magnitude. The states in Model 2 need to be transformed to avoid numerical instability and to be able to implement the model in CSTM. The transformed states in Model 2 are identical to Equation (22) with a_1, a_2, b_1 and b_2 as defined in Equation (24).

Model 3 Because of the risk that the uncertainty intervals might become too large, it was decided to investigate a reduced state dependency and introduce a γ_i parameter. More specifically, Model 3 is expressed with the diffusion matrix

$$\boldsymbol{\sigma}(\boldsymbol{S}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta}) = \begin{bmatrix} \sigma_1 S_{1,t}^{\gamma_1} & 0 \\ 0 & \sigma_2 S_{2,t}^{\gamma_2} \end{bmatrix}.$$

Here, the ith diffusion term is assumed to be dependent on the ith state to the power of γ_i. The Lamperti transform is also required for Model 3 since the diffusion is state dependent. The transformation is identical to Equation (20) with a_1, a_2, b_1 and b_2 as defined in Equation (24).

4 Results

4.1 Searching for optimal γ_i parameters in Model 3

Because of instability related problems with estimating the γ_i parameters in Model 3, an iterative approach had to be adopted to pinpoint the optimal γ_i parameters. Repeatedly the γ_i parameters were adjusted and the corresponding log-likelihood value calculated in search of the maximum log-likelihood area. Figure 4 displays the resulting surface for the profile log-likelihood, varying with the two diffusion parameters γ_1 and γ_2.

Figure 4 shows that an increase for γ_2 causes a linear increase in the log-likelihood, implying that optimal diffusion parameter γ_2 is one. A similar linear correspondence appears between the values of γ_1 and the log-likelihood, but for higher values of the parameter the contour lines even out, meaning that a rather minor increase in the log-likelihood is obtained for further increases in γ_1. It should be recalled that γ_1 is important for controlling the variance of the modelled flow during rain because most of the uncertainty is expected to origin from an insufficient rain input. However, the argument for introducing the γ_i parameters in the first place was to downsize the uncertainty boundaries which might be important when a prediction horizon of more than one step is needed. Therefore, to test the influence on the uncertainty bounds (in this paper only on the one step prediction) $(\gamma_1, \gamma_2) = (0.6, 0.95)$ was selected for further analysis with Model 3.

4.2 Estimation results

Table 2 displays the mean and standard deviation of the estimated parameters. Considering the runoff parameters of the drift term, (A, K and α), it is noticed that the model parameters differ considerably, particularly between Model 1 and Models 2-3, even though the models differ solely with respect to the diffusion term. The drift term of the model remains the same in all three models but the estimated drift term parameters still differ. This shows the importance of selecting a proper description of the diffusion term. The size of the contributing catchment area A is estimated in the range of 35-51 ha, the retention time

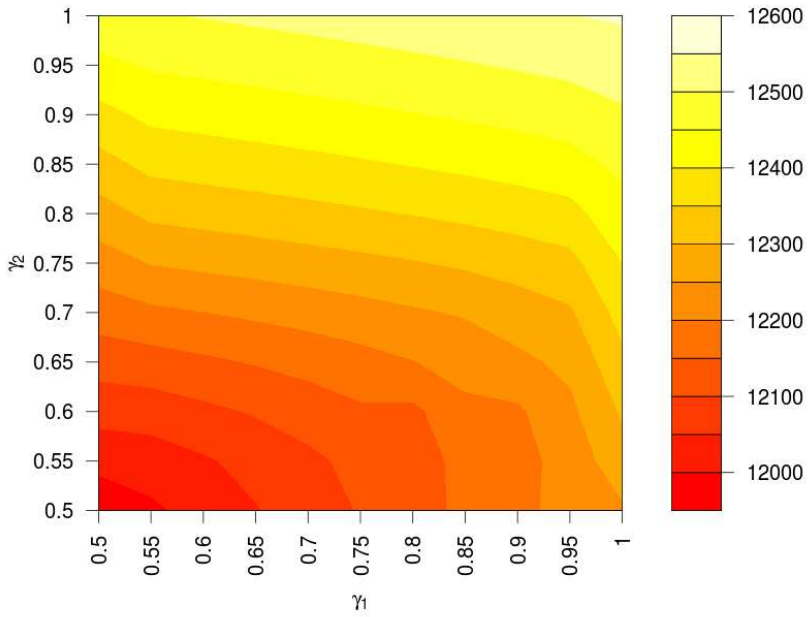

Figure 4: Contour plot of the Log-likelihood as a function of the two diffusion parameters γ_1 and γ_2.

K in the range of 3-5.3 hours and the rain gauge weighting parameter α range

Table 2: Estimation Results

Parameter	Model 1		Model 2		Model 3	
	$\hat{\theta}_{M1}$	sd($\hat{\theta}_{M1}$)	$\hat{\theta}_{M2}$	sd($\hat{\theta}_{M2}$)	$\hat{\theta}_{M3}$	sd($\hat{\theta}_{M3}$)
s_1	-46.641	5.288	-65.645	2.876	-63.580	2.824
c_1	-96.282	5.089	-51.814	3.564	-56.725	2.386
s_2	-48.185	3.528	-35.459	1.882	-39.047	1.544
c_2	17.934	3.726	17.576	1.926	18.039	1.829
$\log(A)$	3.567	0.035	3.934	0.060	3.856	0.064
α	0.398	0.056	0.305	0.081	0.269	0.034
a_0	314.290	4.172	317.330	5.002	308.890	4.154
K	2.999	0.068	5.286	0.220	5.261	0.202
$\log(\sigma_1)$	5.240	0.031	-1.414	0.052	1.107	0.050
$\log(\sigma_2)$	3.053	0.072	-2.444	0.011	-2.082	0.010
$\log(S)$	-7.519	0.047	-19.020	11.559	-19.070	8.845

between 0.3-0.4, that is to say rain gauge P_{321} represents most of the runoff. This is a little surprising since P_{316} is located much closer to the largest paved area of the catchment (cf. Figure 2). Considering the estimated wastewater parameters, (a_0,s_1,s_2,c_1,c_2), it is noticeable that all models returned similar values for a_0 and c_2, whereas the rest of the parameters differ.

Turning to the diffusion parameters, it is seen that for all three models σ_1 are larger than σ_2, which is reasonable since the input uncertainty primarily can be assigned to σ_1. However, the model structure limitations can probably be equally attributed to both states and, thus, a significant σ_2 is found in all three models. The estimated diffusion parameters of the three models cannot be directly compared because in Model 1 the diffusion parameters are constants, whereas for Model 2 the diffusion parameters are scaled with the states and, for Model 3, the state to the power of γ_i. This explains why a decrease of their values are realised with increasing state dependency. The variance of the observation noise S is significant for Model 1 and insignificant for Models 2 and 3. This indicates that the state dependent models cannot separate uncertainty that originate from input and model structural errors from uncertainty that origins from flow measurement errors.

4.3 Model comparison and validation

Table 3 shows that for the one step ahead prediction, Model 2 gives the best fit and uncertainty description according to the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). This means that the state proportional scaling of the diffusion parameters is the preferred diffusion term, although the scaling of the prediction bounds might become a problem if several prediction steps are needed. This is however not investigated in this paper but will be examined in a future study.

Model validation is only considered for the best model (Model 2). A structural behaviour in the residuals would suggest that more physics is needed in the drift term. Figure 5 displays the results of the residual analysis. From the standardised residual plot of Model 2 shown in Figure 5a it seems that the

Table 3: Model Comparison

	$\log(L)$	DF	AIC	BIC
Model 1	11379.81	13	-22733.62	-22643.12
Model 2	12555.67	13	-25085.34	-24994.84
Model 3	12461.81	15	-24893.62	-24789.19

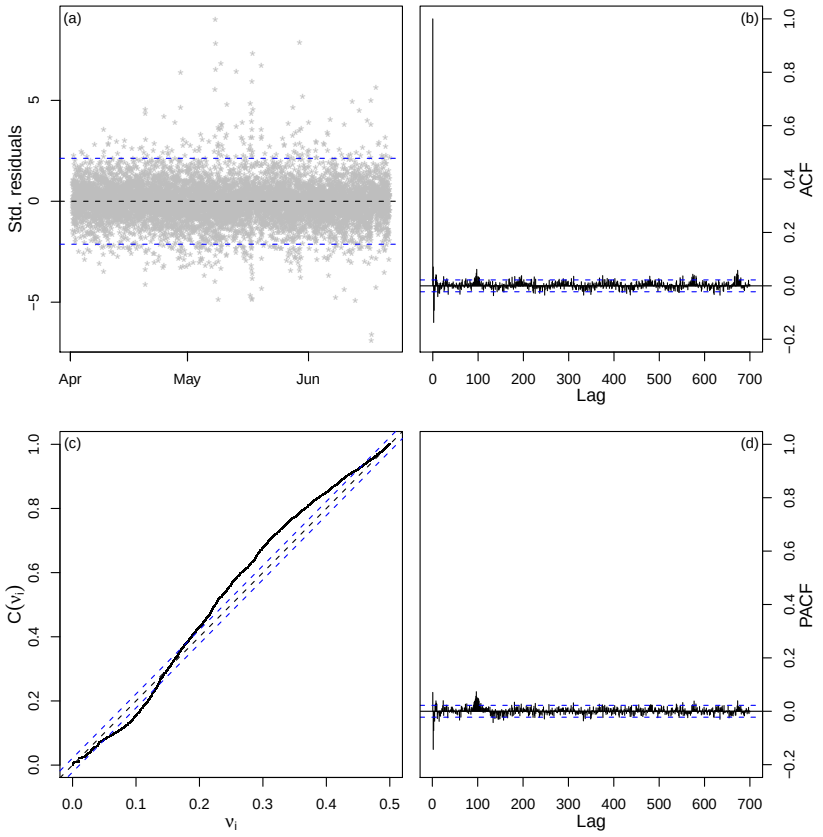

Figure 5: Model validation. (a): Standardised residual plot; (b): Autocorrelation function (ACF) (c): Cumulative periodogram; (d): Partial autocorrelation function (PACF).

Gaussian assumption is satisfied, since the residuals are randomly distributed around zero. Even though few data points appear to depart from the assumption they are not considered to violate the Gaussianity.

Inspecting the autocorrelation functions in Figure 5b and Figure 5d, a minor significance for lags 2 and 3 is visible, but considered small enough to be neglected. However, it is also noticed from the ACF and PACF plots that there is a periodicity in the residual series, note the peaks around lag 96 and 672 corresponding to one day and one week, respectively. These values are also very small and thus ignored here, though it points to a need for further model development of the dry weather flow parameterisation. In the adopted modelling approach, no distinction between weekends (holidays) and working days or

between consecutive working days was tested, although the wastewater diurnal pattern changes accordingly. Thus, the periodicity would be a good starting point to improve the dry weather part of the model, but this is beyond the scope of this paper.

The cumulative periodogram for the residuals is shown in Figure 5c. For the residuals to be considered white, the black solid line should be close to the dashed diagonal line and within the two off-diagonal dashed lines, which correspond to 95% confidence limits for the assumed Gaussianity. In the plot a minor periodicity is detected on each side of the straight line, but these effects are rather limited and can be ignored.

To sum up; the minor deviation for the residuals from the Gaussian assumption for the residuals does not give solid basis for model rejection and Model 2 can be considered sufficiently accurate for assessing the one step prediction uncertainties.

4.4 State and flow uncertainty in dry and wet weather periods

In Figure 6 a comparison of the 95% one step ahead prediction interval of the states is shown for a large rain event, (left column), and a dry weather period, (right column). Notice the scale difference of the vertical axis. For Model 1 the prediction interval of the states remains constant in dry and wet weather and at one point encloses negative state volumes in dry weather. This shows why a state dependent diffusion term is needed. Furthermore, it is clearly seen that the prediction interval is wider for S_1 than S_2 which is related to the uncertain rain input that primarily influences S_1. The prediction interval of Models 2 and 3 reveals that the lower boundary stays positive in dry weather and that the uncertainty increases considerably with the state magnitude, but as expected less in Model 3 than Model 2. Generally, much more water is stored in the states of Model 2 and 3 than was the case with Model 1. This is reasonable since the estimated catchment area is larger for Models 2 and 3 than for Model 1. Moreover, the estimated retention time in Models 2 and 3 is also larger, i.e. in order to obtain the same average dry weather flow, a larger amount of water is stored in both states.

In the left column of Figure 7 the observed flow rate and the corresponding one step ahead 95% prediction interval are displayed for all three models during a rain event and in the right column during a dry weather period. The prediction interval for Model 1 is seen to increase with flow magnitude, which is a consequence of scaling the variance of the observation noise S with the observation function $\boldsymbol{h}$. The prediction interval of Model 1 is the most narrow for large flows, while the opposite holds in dry weather periods. Comparing Model 2

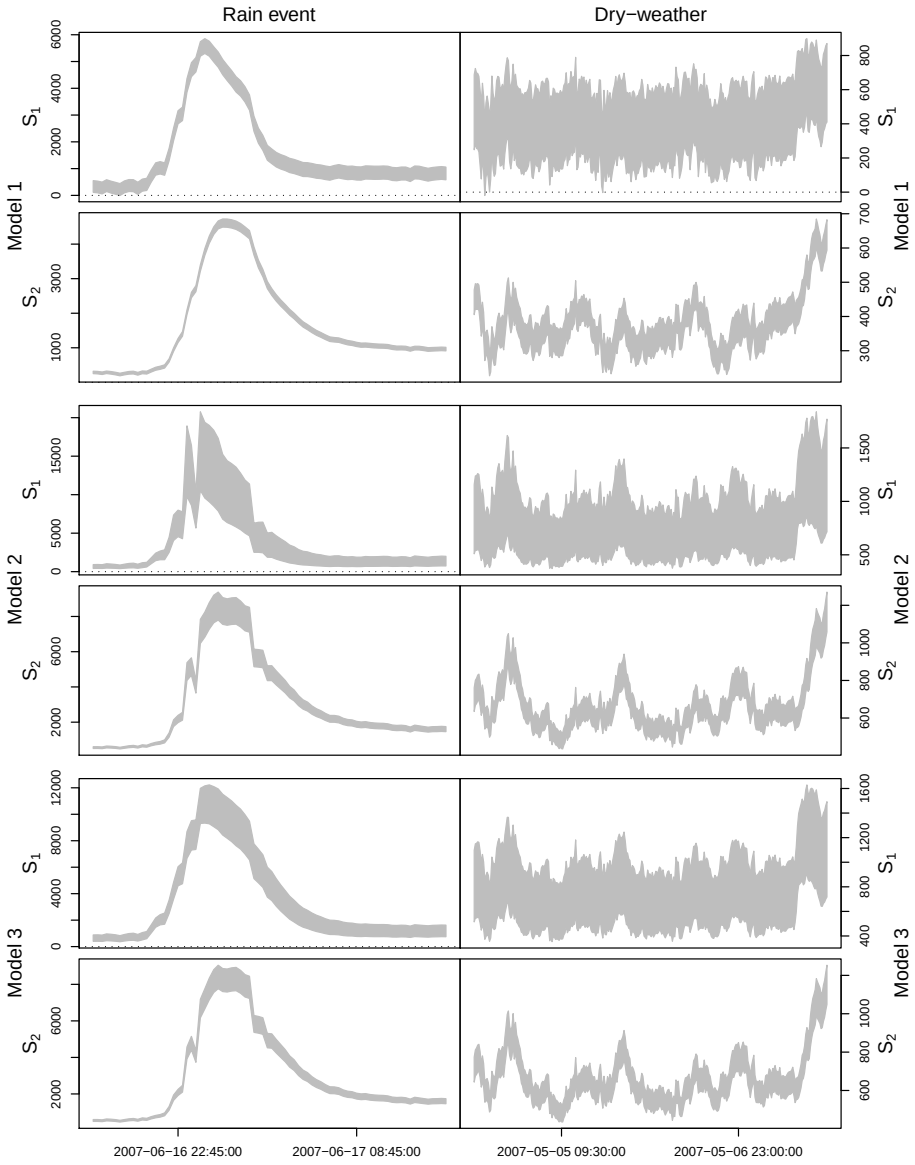

Figure 6: 95% state prediction intervals for all three considered models. State predictions during a rain event is displayed in the left column and in dry-weather in the right column.

with Model 3, the downsizing of the prediction interval is only recognised at

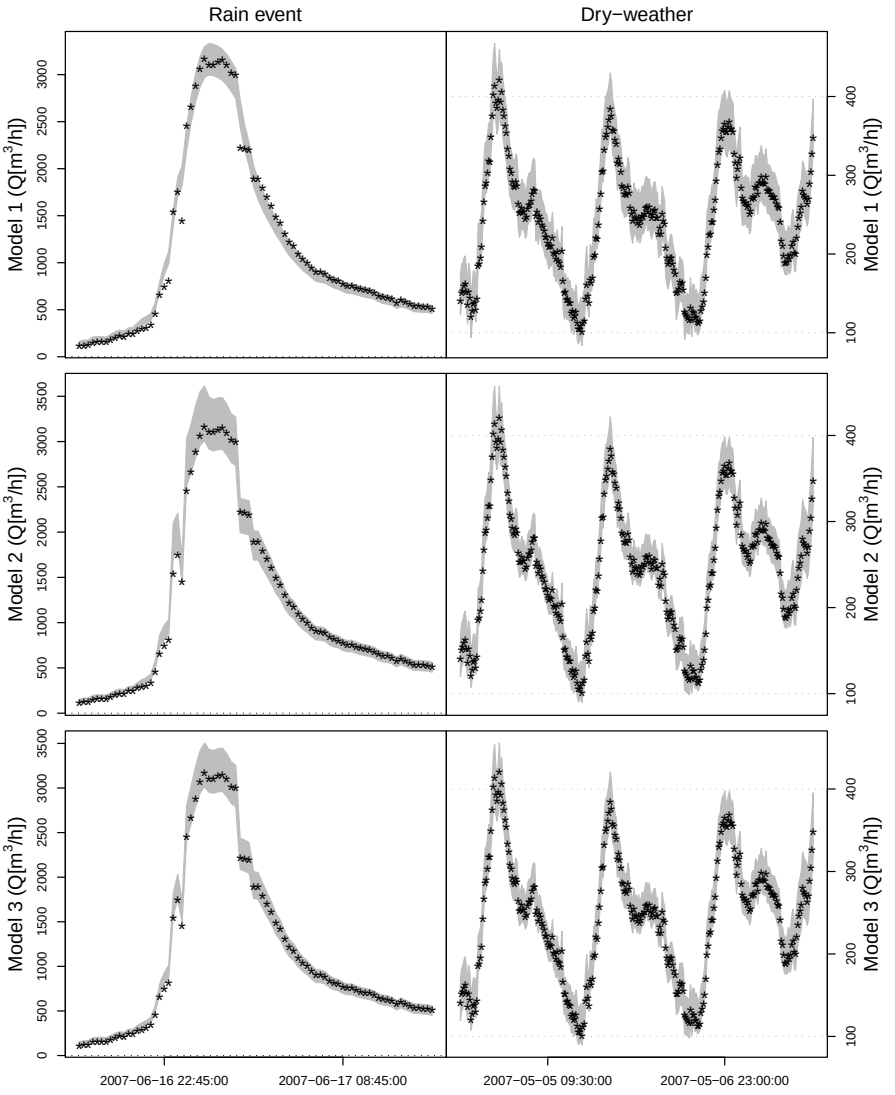

Figure 7: 95% flow prediction intervals (grey area) during wet-weather (left column) and dry-weather (right column) conditions for all three considered models. Measured values are displayed as stars.

the flow peak during rain. However, a longer prediction horizon would prob-

ably lead to a more substantial difference.

Considering how the model assimilates the observations, it can be shown that the observation noise plays an important role. In Model 1 the belief in the drift term of the model is quite good as the updating of the states in the model is not overly aggressive. The predictions are clearly not tracking the latest observation whereas, in the case of Model 2 and Model 3 the states are updated in accordance with the latest observation because observations are taken to be almost 100% accurate. The problem with identifying the observation noise is probably related to both inadequate rain inputs, as well as periods with poor or erroneous flow meter observations.

5 Conclusions

This study has shown that a simple grey box model consisting of two linear reservoirs for rainfall-runoff flow and a harmonic function for wastewater flow can be successfully applied to model the one step prediction uncertainty when an appropriate diffusion term is identified. Such a simple model is attractive for forecasting and control. Three different models were compared that differed with respect to the diffusion term formulation only: one with additive diffusion, one with state proportional diffusion and one with state exponentiated diffusion. To implement the state dependent transformations, it was necessary to apply Itô's formula and the Lamperti transformation. The state proportional diffusion was found to best and adequately describe the one step flow prediction uncertainty while the additive diffusion term resulted in a violation of the physical constraints of the model states that are positively restricted. In a similar manner the risk of obtaining negative flows from an additive observation noise description was avoided by a logarithmic transformation of the observations. This ensured that the observation noise was scaled with the model output. Finally it was found that a proper description of the diffusion term is important for estimation of the physical parameters.

Acknowledgements

We appreciate the help and support of flow meter data from Spildevandscenter Avedøre I/S, and the help with the graphical layout by Lisbeth Brusendorf, DTU Environment. This research project was financially supported by a PhD fellowship co-funded by Krüger A/S, DTU Environment and the Ministry of Science, Technology and Innovation through the graduate school for Urban

Water Technology (UWT), and by the Danish Council for Strategic Research (SWI).

References

Baadsgaard M, Nielsen JN, Spliid H, Madsen H, Preisel M (1997) Estimation in stochastic differential equations with a state dependent diffusion term, *SYSID '97 - 11th IFAC symposium of system identification, IFAC*.

Barbera PL, Lanza LG, Stagi L (2002) Tipping bucket mechanical errors and their influence on rainfall statistics and extremes, *Water Science and Technology* **45**(2):1–9.

Bechmann H, Nielsen MK, Madsen H, Poulsen NK (1999) Grey-box modelling of pollutant loads from a sewer system, *Urban Water* **1**:71–78.

Bechmann H, Madsen H, Poulsen NK, Nielsen MK (2000) Grey box modeling of first flush and incoming wastewater at a wastewater treatment plant, *Environmetrics* **11**:1–12.

Bertrand-Krajewski JL, Bardin JP, Mourad M, Béranger Y (2003) Accounting for sensor calibration, data validation, measurement and sampling uncertainties in monitoring urban drainage systems, *Water Science and Technology* **47**(2):95–102.

Carstensen J, Nielsen MK, Strandbæk H (1998) Prediction of hydraulic load for urban strom control of a municipal WWT plant, *Water Science and Technology* **37**(12):363–370.

Deletic A, Dotto CBS, McCarthy DT, Kleidorfer M, Freni G, Mannina G, Uhl M, Henrichs M, Fletcher TD, Rauch W, Bertrand-Krajewski JL, Tait S (2011) Assessing uncertainties in urban drainage models, *Physics and Chemistry of the Earth* Parts A/B/C, In Press.

El-Din AG, Schmith DW (2002) A neural network model to predict the wastewater inflow incorporating rainfall events, *Water Research* **36**(5):1115–1126, DOI:10.1016/S0043-1354(01)00287-1.

Freni G, Mannina G (2010) Uncertainty in water quality modelling: The applicability of Variance Decomposition Approach, *Journal of Hydrology* **394**(3-4):324–333.

Gelfan A, Hajda P, Novotny V (1999) Recursive system identification for real-time sewer flow forecasting, *Journal of Hydrologic Engineering* **4**(3):280–287.

Giraldo JM, Leirensa S, Díaz-Granados MA, Rodríguez J (2010) Nonlinear optimization for improving the operation of sewer systems: the Bogotá case

study, International Environmental Modelling and Software Society (iEMSs), *2010 International Congress on Environmental Modelling and Software*.

Harremoës P, Madsen H (1999) Fiction and reality in the modelling world-balance between simplicity and complexity, calibration and identifiability, verification and falsification, *Water Science and Technology* **39**(9):1–8.

Iacus SM (2008), *Simulation and Inference for Stochastic Differential Equations - with R Examples*, Springer series of Statistics.

Jacobsen JL, Madsen H (1996) Grey box modelling of oxygen levels in a small stream, *Environmetrics* **7**:109–121.

Jacobsen JL, Madsen H, Harremoës P (1997) A stochastic model for two-station hydraulics exhibiting transient impact, *Water Science and Technology* **36**(5):19–26.

Jazwinski AH (2007) *Stochastic Processes and Filtering Theory*, Dover Publications, Mineola, New York, USA.

Jonsdottir H, Jacobsen JL, Madsen H (2001) A grey-box model describing the hydraulics in a creek, *Environmetrics* **12**:347–356.

Jonsdottir H, Madsen H, Palsson OP (2006) Parameter estimation in stochastic rainfall-runoff models, *Journal of Hydrology* **326**(1-4):379–393.

Jonsdottir H, Nielsen HA, Madsen H, Eliasson J, Palsson OP (2007) Conditional parametric models for storm sewer runoff, *Water Resources Research* **43**:1–9.

Jørgensen HK, Rosenørn S, Madsen H, Mikkelsen PS (1998) Quality control of rain data used for urban runoff systems, *Water Science and Technology* **37**(11):113–120.

Kleidorfer M, Deletic A, Fletcher TD, Rauch W (2009) Impact of input data uncertainties on urban stormwater model parameters, *Water Science and Technology* **60**(6):1545–1554, DOI:10.2166/wst.2009.493.

Krämer S, Fuchs L, Verworn HR (2007) Aspects of radar rainfall forecasts and their effectiveness for real time control -the example of the sewer system of the city of vienna, *Water Practice and Technology Software* **2**(2), DOI:10.2166/wpt.2007.042.

Kristensen NR, Madsen H (2003) *Continuous Time Stochastic Modeling - CTSM 2.3 - Mathematics Guide*, Technical Unversity of Denmark.

Kristensen NR, Madsen H, Jørgensen SB (2004a) A method for systematic improvement of stochastic grey-box models, *Computers and Chemical Engineering* **28**(8):1431–1449.

Kristensen NR, Madsen H, Jørgensen SB (2004b) Parameter estimation in stochastic grey-box models, *Automatica* **40**:225–237.

Kuczera G, Kavetski D, Franks S, Thyer M (2006) Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters, *Journal of Hydrology* **331**(1-2):161–177, DOI:10.1016/j.jhydrol.2006.05.010.

Lei JH, Schilling W (1996) Preliminary uncertainty analysis - a prerequisite for assessing the predictive uncertainty of hydrologic models, *Water Science and Technology* **33**(2):79–90.

Lindblom E, Madsen H, Mikkelsen PS (2007) Comparative uncertainty analysis of copper loads in stormwater systems using GLUE and grey-box modeling, *Water Science and Technology* **56**(6):11–18.

Luschgy H, Pagés G (2006), Functional quantization of a class of Brownian diffsuions: A constructive approach, *Stochastic Processes and their Applications* **116**:310–336.

Madsen H (2008) *Time Series Analysis*, Chapman & Hall/CRC.

Mannina G, Freni G, Viviani G, Saegrov S, Hafskjold LS (2006) Integrated urban water modelling with uncertainty analysis, *Water Science and Technology* **54**(6-7):379–386, DOI:10.2166/wst.2006.611.

Molini A, Lanza LG, Barbera PL (2005) The impact of tipping-bucket rain gauge measurement errors on design rainfall for urban-scale applications, *Hydrological processes* **19**:1073–1088, DOI:10.1002/hyp.5646.

Møller JK, Madsen H (2010) From state dependent diffusion to constant diffusion in stochastic differential equations by the Lamperti transform, Tech. rep., DTU Informatics.

Nielsen HA, Madsen H (2006) Modelling the heat consumption in district heating systems using a grey-box approach, *Energy and Buildings* **38**(1):63–71.

Ocampo-Martinez C, Puig V (2009) On modelling approaches for receding-horizon control design applied to large-scale sewage systems, *Joint 48th IEEE Conference on Decision and Control and 28th Chinese Control Conference* Shanghai, PR China, December 16-18, pp 8052–8058.

Øksendal B (2003), *Stochastic differential equations - an introduction with applications* (6th ed.), Springer.

Pedersen L, Jensen NE, Christensen LE, Madsen H (2010) Quantification of the spatial variability of rainfall based on a dense network of rain gauges, *Atmospheric Research* **95**(4):441–454, DOI:10.1016/j.atmosres.2009.11.007.

Priestley MB (1981) *Spectral Analysis and Time Series*, Series of Monographs and Textbooks, Academic Press, London.

Puig V, Cembrano G, Romera J, Quevedo J, Aznar B, Ramón G, Cabot J (2009) Predictive optimal control of sewer networks using coral tool: application to Riera Blanca catchment in Barcelona, *Water Science and Technology* **60**(4):347–354.

Shedekar VS, King KW, Brown LC, Fausey NR, Heckel M, Harmel RD (2009) Measurement errors in tipping bucket rain gauges under different rainfall intensities and their implication to hydrologic models, *ASABE Annual International Meeting*, June 21-24 pp 1–9.

Tan PC, Berger CS, Dabke KP, Mein RG (1991) Recursive identification and adaptive prediction of wastewater flows, *Automatica* **27**(5):761–768.

Tornøe CW, Jacobsen J, Pedersen O, Hansen T, Madsen H (2004) Grey-box modelling of pharmacokinetic/pharmacodynamic systems, *Journal of Pharmacokinetics and Pharmacodynamics* **31**(5):401–417.

Vaes G, Willems P, Berlamont J (2005) Areal rainfall correction coefficients for small urban catchments, *Atmospheric Research* **77**(1-4):48–59, DOI: 10.1016/j.atmosres.2004.10.015.

Vestergaard M (1998) *Nonlinear filtering in stochastic volatility models*, Master thesis, Technical University of Denmark, Department of Mathematical Modelling, Lyngby, Denmark.

Willems P (2001) Stochastic description of the rainfall input errors in lumped hydrological models, *Stochastic Environmental Research and Risk Assessment* **15**:132–152, DOI: 10.1007/s004770000063.

Willems P (2010) Parsimonious model for combined sewer overflow pollution, *Journal of Environmental Engineering* **136**(3):316–325, DOI:10.1061/(ASCE)EE.1943-7870.0000151.

Willems P, Berlamont J (2002) Probabilistic emission and immission modelling: case-study of the combined sewer-wwtp-receiving water system at Dessel (Belgium), *Water Science and Technology* **45**(3):117–124.

PAPER F

Evaluation of probabilistic flow predictions in sewer systems using grey box models and a skill score criterion

Authors:
F. Ö. Thordarson,A. Breinholt, J. K. Møller, P. S. Mikkelsen,
M. Grum, H. Madsen

Accepted:
Stochastic Environmental Research and Risk Assessment (2012)

Evaluation of probabilistic flow predictions in sewer systems using grey box models and a skill score criterion

Fannar Örn Thordarson[1], Anders Breinholt[2], Jan Kloppenborg Møller[1], Peter Steen Mikkelsen[2], Morten Grum[3], Henrik Madsen[1]

Abstract

In this paper we show how the grey box methodology can be applied to find models that can describe the flow prediction uncertainty in a sewer system where rain data are used as input, and flow measurements are used for calibration and updating model states. Grey box models are composed of a drift term and a diffusion term, respectively accounting for the deterministic and stochastic part of the models. Furthermore, a distinction is made between the process noise and the observation noise. We compare five different model candidates' predictive performances that solely differ with respect to the diffusion term description up to a 4 hour prediction horizon by adopting the prediction performance measures; reliability, sharpness and skill score to pinpoint the preferred model. The prediction performance of a model is reliable if the observed coverage of the prediction intervals corresponds to the nominal coverage of the prediction intervals, i.e. the bias between these coverages should ideally be zero. The sharpness is a measure of the distance between the lower and upper prediction limits, and skill score criterion makes it possible to pinpoint the preferred model by taking into account both reliability and sharpness. In this paper, we illustrate the power of the introduced grey box methodology and the probabilistic performance measures in an urban drainage context.

Key words:
Grey box modelling, Interval prediction, Reliability, Sharpness, Skill score, Urban drainage

[1]Informatics and Mathematical Modelling, Bldg. 305 DTU, DK-2800 Kgs. Lyngby, Denmark
[2]Department of Environmental Engineering, Bldg. 113 DTU, DK-2800 kgs. Lyngby, Denmark
[3]Krüger, Veolia Water Solutions and Technologies, Gladsaxevej 363, DK-2860 Søborg, Denmark

1 Introduction

Sewer flow predictions can, in combination with Model Predictive Control (MPC), be used to minimise damages in a broad sense, e.g. to reduce combined sewer overflows to prevent sludge escaping from wastewater treatment plants and to avoid flooding of vulnerable urban areas. To the authors knowledge, most, if not all, the suggested MPC solutions that have been proposed in the literature to date are based on deterministic models, (see e.g. *Ocampo-Martinez and Puig*, 2010, *Puig et al.*, 2009, *Giraldo et al.*, 2010), even though it is commonly accepted that large uncertainties are present in simulation and prediction with urban drainage models due to unreliable level or flow meters (*Bertrand-Krajewski et al.*, 2003), non-representative rainfall inputs (*Pedersen et al.*, 2010, *Vaes et al.*, 2005, *Willems*, 2001) and/or unreliable rain gauge measurements (*Barbera et al.*, 2002, *Molini et al.*, 2005, *Shedekar et al.*, 2009).

For urban drainage systems, we are still awaiting this shift of paradigm from deterministic to stochastic models in predictive control. This can most likely be attributed to inadequate measurement collection, both with respect to rainfall monitoring/forecasting and in-sewer flow or level metering. However, as the number of measurement devices increase and these devices become more accurate, the potential for building suitable stochastic models also improves. A necessary first step is to derive stochastic models that can describe the predictive uncertainty sufficiently well for a certain prediction horizon of interest. Another important step is to set up a prediction performance evaluation method to be able to compare the predictive performance of different model candidates. In this paper we intend to take these necessary first steps by considering a case catchment area from where both rainfall and flow meter measurements are available for stochastic model building and prediction evaluation of sewer flows.

We apply the grey box methodology as introduced by *Kristensen et al.* (2004a). The grey box approach is based on a state space model where the dynamics are described using Stochastic Differential Equations (SDEs), which contain a drift term and a diffusion term. The grey box methodology has been successfully applied in numerous fields for stochastic model building, including e.g. pharmacology (*Tornøe*, 2004), chemical engineering (*Kristensen et al.*, 2004a,b), district heating (*Nielsen and Madsen*, 2006), hydrology (*Jonsdottir et al.*, 2001, 2006) and ecology (*Møller et al.*, 2011). We give particular attention to the diffusion term by considering various diffusion term descriptions. Several tools have been developed to validate and compare models, especially for point forecasts that exclusively rely on the single value prediction. In contrast, little attention has been given to interval predictions, which play a crucial role in stochastic control design. We propose here to use a skill scoring criterion for interval

prediction evaluation, and show how this can be applied to find the preferred model among the candidate models for a specific prediction horizon. The skill scoring criterion has previously been applied for prediction evaluation purposes in wind power generation (see *Pinson et al.*, 2007, *Møller et al.*, 2008).

In Section 2, we outline the stochastic grey box methodology. Section 3 includes a description of the interval prediction generation and how the prediction performance can be evaluated on the basis of the reliability, the sharpness and the skill score criterion. Section 4 illustrates the applicability of the grey box methodology and the use of the prediction performance criteria as important tools for model selection. Finally, in section 5 we conclude on our findings.

2 The stochastic grey box model

2.1 Model structure

The model used in this study is a grey box model, or a continuous-discrete time stochastic state space model, represented by

$$d\boldsymbol{X}_t = \boldsymbol{f}(\boldsymbol{X}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta})dt + \boldsymbol{\sigma}(\boldsymbol{X}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta})d\boldsymbol{\omega}_t \tag{1}$$

$$\boldsymbol{Y}_k = \boldsymbol{g}(\boldsymbol{X}_k, \boldsymbol{u}_k, t_k, \boldsymbol{\theta}) + \boldsymbol{e}_k. \tag{2}$$

where the first equation is called the system equation, composed of a set of SDEs in continuous time. The states are partially observed in discrete time through the observation equation (2). The time is $t \in \mathbb{R}_0$ and t_k (for $k = 1, \ldots, K$) are the discretely observed sampling instants for the K available measurements. The states in the system equation $\boldsymbol{X}_t \in \mathbb{R}^n$ describe the system dynamics in continuous time, whereas $\boldsymbol{X}_k \in \mathbb{R}^n$ in the observation equation is the observed states in the discrete time as specified by the observations. The input variables are represented by the vector $\boldsymbol{u}_t \in \mathbb{R}^m$ and the vector of the measured output variables $\boldsymbol{Y}_k \in \mathbb{R}^l$. The vector $\boldsymbol{\theta} \in \mathbb{R}^p$ includes the unknown parameters that characterise the model, and the functions $\boldsymbol{f}(\cdot) \in \mathbb{R}^n$, $\boldsymbol{\sigma}(\cdot) \in \mathbb{R}^{n \times n}$ and $\boldsymbol{g}(\cdot) \in \mathbb{R}^l$ form the structural behaviour of the model. The measurement error $\boldsymbol{e}_k$ is assumed to be a l-dimensional white noise process with $\boldsymbol{e}_k \sim N(\boldsymbol{0}, \boldsymbol{V}(\boldsymbol{u}_k, t_k, \boldsymbol{\theta}))$, where $\boldsymbol{V}$ is the covariance of the measurement error, and $\boldsymbol{\omega}_t$ is a n-dimensional standard Wiener process. The first term in the system equation is the drift term, representing the dynamic structure of the system that is formulated by ordinary differential equations. The second term is the diffusion term which corresponds to the process noise related to the particular state variable in the state-space formulation.

Discrepancies between output from deterministic models and measurements are often referred to as measurement errors, even though the consecutive residuals are clearly auto-correlated. In reality, these auto-correlated discrepancies can however be explained by both non-representative and/or faulty inputs as well as model structural deficiencies. Consequently, a distinction between measurement noise and noise related to inputs and model deficiencies is required. The stochastic grey box model provides such a distinction by separating the process noise from the output measurement noise, where the process noise as described by the diffusion term is related to the state variables and accounts for noise that is not related to the output measurements.

2.2 Parameter estimation and state transformation

For parameter estimation the Maximum Likelihood (ML) method is used, and the Kalman Filter techniques are applied to evaluate the likelihood function (*Jazwinski*, 2007). For the grey box model in equations (1) and (2), the unknown model parameters are obtained by maximising a likelihood function that is a product of the one-step conditional densities (*Madsen*, 2008). Hence, the estimated parameters for an adequate model correspond to a fit where the distribution for the residual series for the one-step ahead prediction error is assumed to be serial independent and Gaussian. However, utilising such a model for predictions covering more than one-step ahead usually results in a residual series that is correlated, and when dealing with increasing prediction horizon, the predictive distribution for the output may divert from the assumed normality.

To estimate the unknown parameters of the model, the software CTSM[1] (*Kristensen and Madsen*, 2003) is used. The software is well suited for estimation of linear and many nonlinear systems. In CTSM, the ordinary Kalman filter gives the exact solution for the state estimation for linear systems, whereas the extended Kalman filter provides an approximation for the states for nonlinear systems.

Parameter and state estimation is not possible with CTSM if state dependency is included in the diffusion term, as this requires higher order filtering techniques to solve the estimation than are available in the extended Kalman filter techniques implemented in the software (*Vestergaard*, 1998). However, efficient and numerically stable estimates can be obtained by considering a transformation of the states. In particular, the transformation is well-suited for a SDE when the diffusion term is only dependent on the corresponding state variable. With such a univariate diffusion, it is always possible to transform the

[1] Continuous-Time Stochastic Modelling - www.imm.dtu.dk/ctsm

state description to obtain a state independent diffusion term (*Baadsgaard et al.*, 1997).

The transformation of the ith state variable $X_{i,t}$ to $Z_{i,t}$, for $i = 1, \ldots, n$, is referred to as the Lamperti transform (*Iacus*, 2008) and, subsequently, a corresponding SDE for the transformed variable $Z_{i,t}$, is obtained by Itô's formula (*Øksendal*, 2003). The diffusion in the transformed SDE is state independent and the transformed grey box model is rewritten

$$d\boldsymbol{Z}_t = \widetilde{\boldsymbol{f}}(\boldsymbol{Z}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta})dt + \widetilde{\boldsymbol{\sigma}}(\boldsymbol{u}_t, t, \boldsymbol{\theta})d\boldsymbol{\omega}_t \tag{3}$$

$$\boldsymbol{Y}_k = \widetilde{\boldsymbol{g}}(\boldsymbol{Z}_k, \boldsymbol{u}_k, t_k, \boldsymbol{\theta}) + \boldsymbol{e}_k, \tag{4}$$

where the functions $\boldsymbol{f}(\cdot)$, $\boldsymbol{\sigma}(\cdot)$ and $\boldsymbol{g}(\cdot)$ in Eq's. (1) and (2) have been reformulated, respectively to $\widetilde{\boldsymbol{f}}(\cdot)$, $\widetilde{\boldsymbol{\sigma}}(\cdot)$ and $\widetilde{\boldsymbol{g}}(\cdot)$ in relation to the transformation of the state space. The parameters $\boldsymbol{\theta}$ and the input-output relations are, however, not affected by the transformation.

In this study, it is furthermore anticipated that flow measurement errors increase proportionally with flow magnitude and thus a log-transformation of the observations are needed to secure a Gaussian measurement noise term. This observation transformation results in an observation equation that has an additive noise term (*Limpert et al.*, 2001).

3 Prediction, uncertainty and evaluation

3.1 Uncertainty of h-step ahead prediction

The objective with the proposed grey box model is to predict the sewer flow at time $k + h$, which is denoted as Y_{k+h}. In parallel, we have $\hat{Y}_{k+h|k}$ as the prediction of the flow at time $k + h$, given the available information at time k where h indicates the number of time steps for the prediction. By using the ML method, we find that the optimal prediction is equal to the conditional mean for the model structure (see *Madsen*, 2008). Hence, the prediction is obtained by

$$\hat{\boldsymbol{Y}}_{k+h|k} = E\left[\boldsymbol{Y}_{k+h} | \mathrm{Y}_k, \boldsymbol{u}_{k+h}\right] \tag{5}$$

$$= \widetilde{\boldsymbol{g}}(\hat{\boldsymbol{Z}}_{k+h|k}, \boldsymbol{u}_{k+h}, t_{k+h}, \boldsymbol{\theta}), \tag{6}$$

meaning that for a given sequence of precipitation input up to time $k + h$ and observed flow up to time k, $\mathrm{Y}_k = [\boldsymbol{Y}_k, \ldots, \boldsymbol{Y}_0]^\top$, the state prediction at time $k + h$

can be estimated and consequently supply the observation equation with a suitable description for the prediction. The challenge in predicting the future flow in the system is then not directly related to predictions based on the observation equation, but rather on predicting the state variables in the system equation. The state prediction can be accomplished by considering the conditional expectation of the future state:

$$\hat{\boldsymbol{Z}}_{k+h|k} = E\left[\boldsymbol{Z}_{k+h}|\hat{\mathrm{Y}}_k, \boldsymbol{u}_{k+h}\right], \tag{7}$$

i.e. the conditional mean of $\boldsymbol{Z}_{k+h}$ given all measurements up to time k (*Madsen*, 2008).

In the following study, the grey box model in equations (1) and (2) is used to describe the model structure, whereas the transformed model is used for parameter estimation and model prediction in equations (3) and (4). As mentioned in section 2.2, the Gaussian assumption for the model output is only valid for one-step ahead predictions. Thus for $h \geq 1$, a numerical approach is considered, i.e. an Euler scheme for the SDEs in the system equation (3) is applied to predict the sewer runoff (*Kloeden and Platen*, 1999). Thus, a sufficient probability distribution for the h-step ahead prediction is obtained by generating a number of simulations from each time step, and from this empirical distributions can be derived for the prediction intervals.

3.2 Prediction intervals

The ideal coverage of the prediction interval is defined as the nominal coverage $1-\beta$, $\beta \in [0,1]$. The upper and lower limits of the interval prediction are obtained from quantile forecasts, which are easy to obtain with a large number of simulations provided for the same prediction horizon, resulting in a reasonable empirical probability distribution for the sewer flow. If $F_{k+h|k}$ is the cumulative distribution function of the random variable $\hat{Y}_{k+h|k}$ and $\tau \in [0,1]$ is the proportion of the relative quantile, the τ-quantile forecast for the $k+h$ prediction is obtained by

$$q^{(\tau)}_{k+h|k} = F^{-1}_{k+h|k}(\tau). \tag{8}$$

If $l = \beta/2$ and $u = 1-\beta/2$ are defined as the lower and upper quantiles for the prediction interval at level $1-\beta$, respectively, the prediction interval for the lead time $k+h$, issued at time k, can be described as

$$\hat{I}^{(\beta)}_{k+h|k} = \left[\hat{q}^{(l)}_{k+h|k}, \hat{q}^{(u)}_{k+h|k}\right] \tag{9}$$

where $\hat{q}^{(l)}_{k+h|k}$ and $\hat{q}^{(u)}_{k+h|k}$ are, respectively, the lower and upper prediction limits at levels $\beta/2$ and $1-\beta/2$ (*Pinson et al.*, 2007, *Møller et al.*, 2008).

3.3 Reliability

For the prediction interval to be of practical usage for decision makers it is a primary requirement for the interval to be reliable, indicating that the upper and lower limits have to correspond to the nominal coverage rate of $1-\beta$.

To obtain an evaluation of the reliability of the interval we define a counter that rewards prediction intervals that are able to capture the observations. For a given prediction interval, as represented in Eq. (9), and corresponding measured flow in the system Y_{k+h}, the binary indicator variable $n_{k,h}^{(\beta)}$ is obtained by

$$n_{k,h}^{(\beta)} = \begin{cases} 1, & \text{if } Y_{k+h} \in \hat{I}_{k+h|k}^{(\beta)} \text{ for } k \leq K-h \\ 0, & \text{otherwise} \end{cases}, \tag{10}$$

corresponding to hits and misses of the h-step prediction interval. The mean of the binary series then corresponds to the actual proportion of hits in the estimation period, i.e. for prediction horizon h the proportion of hits for a flow series of length K, is given by

$$\bar{n}_h^{(\beta)} = \mathrm{E}\left[n_{k,h}^{(\beta)}\right] = \frac{1}{K-h}\sum_{k=1}^{K-h} n_{k,h}^{(\beta)}. \tag{11}$$

The discrepancy between the nominal coverage and the observed proportion of hits is measured by the bias

$$b_h^{(\beta)} = 1-\beta-\bar{n}_h^{(\beta)}, \tag{12}$$

where a perfect fit is defined as $b_h^{(\beta)} = 0$, i.e. that the empirical coverage is equal to the nominal coverage, $\bar{n}_h^{(\beta)} = 1-\beta$, and a perfect reliability is obtained. However, when the empirical coverage is larger than the nominal, i.e. $\bar{n}_h^{(\beta)} > 1-\beta$, we talk about an overestimation in the coverage. This means that, since the empirical coverage is subtracted from the nominal coverage, we obtain $b_h^{(\beta)} < 0$ when the predictions overestimate the coverage. When the opposite is the case, this is referred to as underestimation, i.e. $b_h^{(\beta)} > 0$.

3.4 Sharpness

Sharpness is an accuracy measure of the prediction interval where smaller values indicate that the model is better suited to generate predictions (*Gneiting et al.*, 2007). The size of the interval prediction, issued at time k for lead time $k+h$ is measured as the difference between the corresponding upper and lower

quantile forecast, and averaging over the whole time series, defines the average sharpness. For the horizon h and coverage $1-\beta$, the sharpness is calculated by

$$\overline{\delta}_h^{(\beta)} = \frac{1}{K}\sum_{k=1}^{K}\left(\hat{q}_{k+h|k}^{(u)} - \hat{q}_{k+h|k}^{(l)}\right) \tag{13}$$

and by calculating $\overline{\delta}_h^{(\beta)}$ at relevant coverages, a δ-diagram can be viewed to summarise the evaluation of the sharpness. When comparing interval predictions generated from different models, the one with the smallest distance between upper and lower bound is the sharpest.

3.5 Interval score criterion and resolution

The skill score combines the performance measures discussed above in a single numerical value, which enables us to compare the predictive performance of different models directly. The skill score for interval predictions is outlined in detail by *Gneiting and Raftery* (2007), where the score of the individual prediction interval is also referred to as an interval score. The skill score Sc for the interval prediction, at time instant k, is calculated as

$$\begin{aligned} Sc_{I,k,h}^{(\beta)} =& (\hat{q}_{k+h|k}^{(u)} - \hat{q}_{k+h|k}^{(l)}) \\ &+ \frac{2}{\beta}(\hat{q}_{k+h|k}^{(l)} - Y_{k+h})\mathbf{1}\{Y_{k+h} < \hat{q}_{k+h|k}^{(l)}\} \\ &+ \frac{2}{\beta}(Y_{k+h} - \hat{q}_{k+h|k}^{(u)})\mathbf{1}\{Y_{k+h} > \hat{q}_{k+h|k}^{(u)}\}, \end{aligned} \tag{14}$$

where the indicator $\mathbf{1}\{\cdot\}$ is equal to one if the inequality within the brackets is fulfilled, but zero otherwise. As the objective is to evaluate the predictive performance of each model by a single number, an extension is required to account for the whole considered period. Hence, we average the scores for all time instants where observations are available, and thus the score becomes independent of the length of the time series. The average interval score criterion for h-step prediction is written

$$\begin{aligned} \overline{Sc}_{I,h}^{(\beta)} =& \frac{1}{K}\sum_{k=1}^{K} Sc_{I,k,h}^{(\beta)} = \bar{\delta}_h^{(\beta)} \\ &+ \frac{2}{\beta(K-h)}\sum_{k=1}^{K-h}\Big[(\hat{q}_{k+h|k}^{(l)} - Y_{k+h})\mathbf{1}\{Y_{k+h} < \hat{q}_{k+h|k}^{(l)}\} \\ &+ (Y_{k+h} - \hat{q}_{k+h|k}^{(u)})\mathbf{1}\{Y_{k+h} > \hat{q}_{k+h|k}^{(u)}\}\Big]. \end{aligned} \tag{15}$$

It follows from Eq. (15) that for any observation that falls outside the predefined prediction interval, the skill score is increased by the distance between the interval and the observation at each considered quantile. Hence, the skill score gives a positive penalisation, which indicates that an increase in the score criterion will result in a reduced fit of the prediction interval. Therefore, we select the prediction interval with the lowest skill score.

The indication of the individual observation in relation to the prediction interval can be merged into an indicator, corresponding to the reliability indicator in Eq. (10). Thus, the interval score in Eq. (15) can be written as an indirect function of the prediction interval in Eq. (9) by including the reliability indicator from Eq. (10), i.e.

$$\begin{aligned}\overline{Sc}_{I,h}^{(\beta)} =& \bar{\delta}_h^{(\beta)} + \frac{2}{\beta(K-h)} \sum_{k=1}^{K-h} \left(1 - n_{k,h}^{(\beta)}\right) \\ & \times \left(\min \left| Y_{k+h} - [\hat{q}_{k+h|k}^{(l)}, \hat{q}_{k+h|k}^{(u)}] \right| \right),\end{aligned} \tag{16}$$

where the second term under the summation accounts for the minimum distance between the observed value and the prediction interval, which is always either the lower or the upper limit of the interval.

The score is still a function of the prediction horizon h. This indicates that there are just as many $\overline{Sc}_{I,h}^{(\beta)}$ as there are h's. To evaluate the performance independently of h, we simply average over all horizons, obtaining the interval score criterion $\overline{\overline{Sc}}_I^{(\beta)}$.

We talk about resolution when conditioning the predictive distributions on some particular property. For urban drainage systems, it is expected that the skill score (or the sharpness and reliability) depends on the weather, i.e. the predictive performance is assumed to be different in periods of dry weather than in periods of wet weather.

4 Application results

In the previous sections, the model framework and tools for assessing the uncertainty and the performance of the model have been described. In the following we introduce the catchment area and the data, the applied grey box models, and finally present and discuss our results.

4.1 Description of the case study

The considered catchment area, which receives both wastewater and rainfall-runoff, is located in the Municipality of Ballerup west of Copenhagen in Denmark; see Figure 1. It is connected to the second largest wastewater treatment plant in Denmark, located in Avedøre. Flow was measured downstream from the catchment area with a semi-mobile ultrasonic Doppler type flow meter. The flow meter was placed in an interceptor pipe with a dimension of 1.4 m. The flow meter logs every 5 minutes, but in this study a temporal resolution of 15 minutes was considered and, thus, only every third available measurement is used.

Precipitation was measured using two tipping bucket gauges with a volumetric resolution of 0.2 mm. The rain gauges are located just outside the considered catchment area, approximately 12 km apart from each other (Fig. 1). Data of flows and rain for almost three month period were used in the case study, i.e. from April 1 2007 to June 21 2007. The considered grey box models were estimated for all three months. However for prediction uncertainty assessment only data from May and June were utilised as very few rain events were logged by the rain gauges in April, and because the rain periods are the most important, it was decided to leave out this month. When generating predictions with the models we used the measured precipitation up to 4 hours ahead of current time assuming a perfect rain forecast was available. This assumption is obviously unrealistic but serves an illustrative purpose here by showing how the skill score terminology can be applied to select the preferred model.

4.2 The stochastic model

The model should be kept simple and identifiable from data to facilitate the parameter estimation. In hydrology it is well known that the rainfall-runoff relationship can often be modelled with a series of linear reservoirs (e.g. *Jacobsen et al.*, 1997, *Mannina et al.*, 2006, *Willems*, 2010). A model with just two reservoirs is considered here, where the volume in each reservoir corresponds to a state variable in the grey box model. There is also a contribution of wastewater from the connected households to the sewer flow that needs to be accounted for. The model is written as

$$\begin{aligned} d\begin{bmatrix} S_{1,t} \\ S_{2,t} \end{bmatrix} &= \begin{bmatrix} \alpha A P_{1,t} + (1-\alpha) A P_{2,t} + a_0 - \frac{2}{K} S_{1,t} \\ \frac{2}{K} S_{1,t} - \frac{2}{K} S_{2,t} \end{bmatrix} dt \\ &+ \begin{bmatrix} \sigma_1 S_{1,t}^{\gamma_1} & 0 \\ 0 & \sigma_2 S_{2,t}^{\gamma_2} \end{bmatrix} d\boldsymbol{\omega}_t, \end{aligned} \tag{17}$$

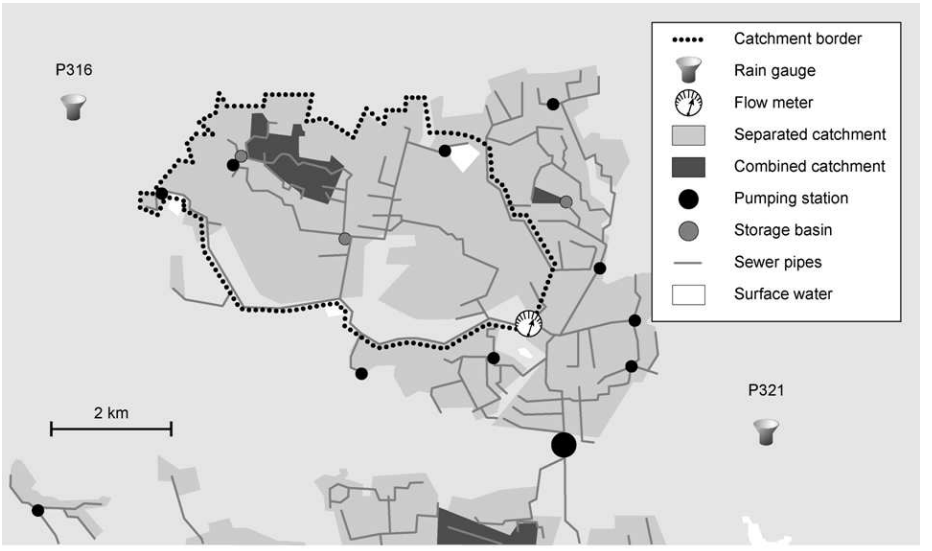

Figure 1: The Ballerup catchment area.

$$\log(Y_k) = \log\left(\frac{2}{K}S_{2,k} + D_k\right) + e_k, \tag{18}$$

where D_k is the wastewater flow variation formulated as a periodic function with diurnal cycles of length L, i.e.

$$D_k = \sum_{i=1}^{2}\left(s_i \sin\frac{i2\pi k}{L} + c_i \cos\frac{i2\pi k}{L}\right) \tag{19}$$

and s_1, s_2, c_1 and c_2 are parameters. The first reservoir $S_{1,t}$ receives runoff from the contributing area A at time t, caused by the rainfall registered at the two rain gauges $P_{1,t}$ and $P_{2,t}$. A weighting parameter α is defined to account for the fraction of the measured runoff that can be attributed to rain gauge $P_{1,t}$, whereas the remaining $1-\alpha$ is attributed to $P_{2,t}$ assuming that the rainfall input area A is fully described by the two rain gauges. The second reservoir, $S_{2,t}$, receives outflow from the first reservoir and diverts it to the flow gauge downstream from the catchment.

To fully account for the wastewater flow in the grey box model, a constant term for the average dry-weather flow a_0 is included. The constant enters the first

state to secure the physical interpretation of the system, i.e. water is always passing through the system, also in dry weather, which means that the reservoirs always contain water. From a modelling point of view this is important because the state variance from the diffusion term in Eq. (17) - if large enough - could lead to predicted states that are negative, which is physically impossible. This risk of receiving negative states is especially high if an additive diffusion term is used and therefore we focus on state dependent diffusion terms only; see *Breinholt et al.* (2011) for more details. When rainwater enters the system, the volume of water in the reservoir increases and the diffusion term is scaled accordingly (see Eq. 17), which means that the state prediction uncertainty rises.

The observation equation (18) depends on the second state variable only, since the output from the second reservoir corresponds to the flow measured downstream from the catchment area. The observation equation is log-transformed to account for proportional observation variance as mentioned in Section 2.2. In the following we will investigate various state dependencies through the γ parameter in each state dependent diffusion in the system equation (17). Different γ parameters will produce different prediction intervals and, subsequently, different skill scores. This is useful for model prediction comparison.

The diffusion parameters γ_1 and γ_2 are restricted to $\gamma_i \in]0.5,1]$, for $i = 1,2$ in the system equation. The reasons are that for $\gamma_i \leq 0.5$ there is a positive probability of reaching zero and the risk of obtaining a non-stationary diffusion process is increased, whilst for $\gamma_i > 1$ the system existence and uniqueness is not guaranteed because the behaviour of the solution might explode in finite time (*Iacus*, 2008).

Five models are proposed with different combinations of the diffusion parameters γ_1 and γ_2. These are $(0.5,0.5)$, $(1,0.5)$, $(0.5,1)$, $(0.75,0.75)$ and $(1,1)$. The minimum γ parameter is actually slightly higher than 0.5 (i.e. 0.5001) in order to fulfill the parameter restriction, but for practical reasons is rounded to 0.5 in the text below. It is not possible to estimate the γ parameters with CTSM because each combination of γ parameters has its own restricted $Z_{i,t}$ domain. To distinguish between the models, they have been designated "M1", "M2", etc., as in the first line in Table 1; the corresponding sets of γ parameters are indicated in the next two rows (highlighted in bold).

4.3 Estimation results

The parameter estimation is shown in Table 1. It is seen that the choice of diffusion term description affects all the parameters to some extent. However, the dry weather parameters s_1,s_2,c_1, c_2 and a_0 are not noticeably influenced, even

though a_0 is slightly higher in M2 than it is in the other models. Considering the wet weather parameters A, K and α, it is seen that A and K are positively correlated with γ_2 while α is estimated to have more or less the same value. The largest area is estimated with M5. Regarding the estimates for the diffusion parameters σ_1 and σ_2, a higher expected parameter value follows a lower state dependency.

4.4 Overall reliability assessment

The average reliability bias is studied in Figure 2, both as a function of the nominal coverage (Fig. 2a), and as a function of the prediction horizon of up to 4h ahead (Fig. 2b). In Figure 2a the reliability bias is calculated as the average for all the considered prediction steps, whereas in Fig. 2b, the reliability bias

Table 1: The results from the parameter estimation, for various values of (γ_1, γ_2), for all five models. Standard deviance is indicated in brackets.

θ	Unit	M1	M2	M3	M4	M5
γ_1	-	**0.500**	**1.000**	**0.500**	**0.750**	**1.000**
γ_2	-	**0.500**	**0.500**	**1.000**	**0.750**	**1.000**
s_1	-	-59.355	-65.313	-63.303	-63.909	-65.545
		(3.927)	(3.861)	(2.764)	(3.310)	(2.709)
s_2	-	-41.363	-34.090	-39.143	-37.341	-34.904
		(2.537)	(3.049)	(1.989)	(2.377)	(2.133)
c_1	-	-61.618	-49.407	-56.898	-51.593	-50.884
		(4.321)	(8.038)	(3.169)	(4.062)	(3.397)
c_2	-	17.437	17.120	18.407	17.133	17.785
		(2.537)	(2.927)	(1.913)	(2.220)	(1.889)
a_0	m^3/h	313.310	345.510	307.000	314.390	319.080
		(4.321)	(1.217)	(4.524)	(5.263)	(5.686)
α	-	0.359	0.374	0.288	0.334	0.335
		(0.068)	(0.080)	(0.070)	(0.059)	(0.067)
A	ha	42.406	39.694	49.591	46.479	51.413
		(1.059)	(1.221)	(1.062)	(1.080)	(1.104)
K	h	4.253	4.104	5.237	4.763	5.221
		(0.148)	(0.472)	(0.200)	(0.201)	(0.274)
σ_1	-	6.510	0.373	5.866	1.313	0.254
		(1.042)	(1.078)	(1.051)	(1.048)	(1.050)
σ_2	-	2.186	1.817	0.087	0.449	0.085
		(1.027)	(1.079)	(1.010)	(1.016)	(1.011)

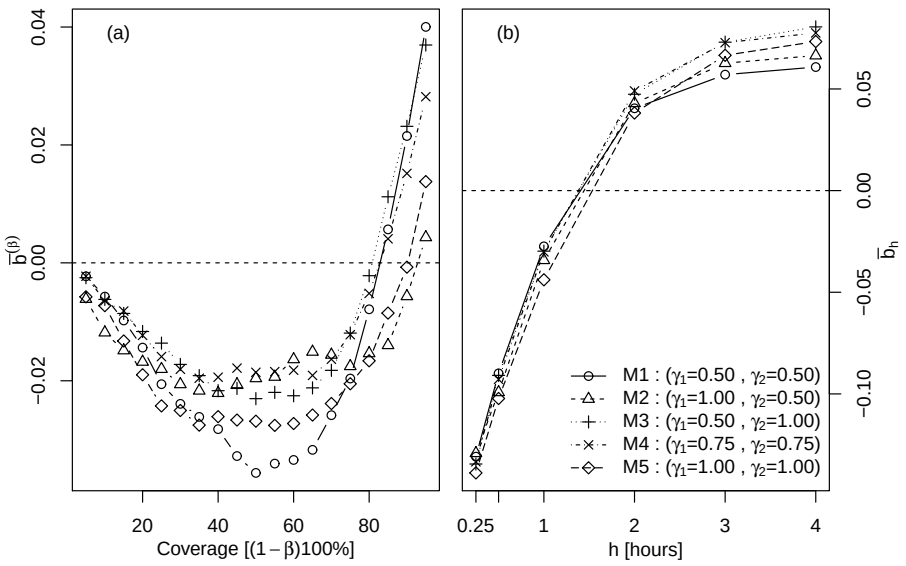

Figure 2: Reliability bias for all five models of interest: (a) averaged over the entire prediction horizon, plotted as a function of the nominal coverage rate, (b) averaged over the coverage rates for each prediction step considered in the study. Coverage rates calculated for the nominal coverage rates:{5%, 10%...95%}.

is calculated as an average of all the nominal coverages. No definite deviation is observed between the models, neither at the chosen prediction steps, nor at different nominal coverages. At coverage up to 80% - 90%, Figure 2a shows that all five models slightly overestimate the nominal coverage, whereas for higher nominal coverage the bias is underestimated. Furthermore, the models approach the nominal coverage at around 85% - 90%.

Regarding the reliability bias for the individual models, Figure 2a reveals that M1 deviates the most from the ideal as it exhibits the largest positive bias at intermediate coverage rates, and the most negative bias at higher nominal coverage rates. M2 is the most reliable model on average, the average bias from ideal reliability is -0.01 for all coverage rates up to 95% coverage. This indicates that (γ_1=1, γ_2=0.5) provides the best reliability across all the considered horizons.

Turning to the average reliability bias as a function of the prediction horizon, Figure 2b shows that all five models produce almost the same reliability structure; i.e. for shorter horizons the reliability bias of the model predictions is overestimated, whereas for horizons longer than 1.5h reliability is increasingly

underestimated. Thus, the almost identical shift from overestimation to underestimation implies that all the models are reliable at 1.5h lead time, but it is recalled that this is an average for all nominal coverages and, thus, it can vary for each nominal coverage. In contrast to what was concluded from Figure 2a, the most reliable model in Figure 2b is M1. However, differences in reliability bias between the models are very small, suggesting that the minor discrepancies for the longer horizons are unimportant.

Here, a single nominal coverage is chosen for further investigation. From the reliability assessment above, it was detected that, on average, the 85% - 90% coverages are reliable. Therefore, the 90% coverage is selected for further investigation, which is also a typical value for interval prediction within hydrology.

4.5 Performance evaluation of the 90% prediction interval

In Figure 3, the reliability bias of the 90% prediction interval ($\beta = 0.1$) as a function of the prediction horizon is seen. The same shift in reliability from overestimation to underestimation is observed for all models as the prediction horizon increases. The deviation from the nominal coverage is generally not that big, although M3 deviates almost 10% at the 4h prediction step. On average, M1 is the most reliable model with mean distance from ideal reliability of 0.043. This can be hard to envisage from Figure 3, because at larger prediction horizons, i.e. more than 1 hour, M1 is clearly less reliable than M2 and M5.

In Figure 4, the sharpness of the 90% prediction intervals is plotted for all the models as a function of the prediction horizon. As expected, all models become less sharp with increasing prediction horizon, i.e. the uncertainty of the prediction rises, but only up to two hours. Hereafter the uncertainty levels out. When considering all prediction horizons, M2 is the least sharp model (the one with the largest uncertainty), and already at the 0.5h prediction step it deviates considerably from the other models. Figure 4 also reveals that the models with $\gamma_1 = 0.5$ prove to be the sharpest for all prediction horizons, and M3 is visually slightly sharper than M1. Thus, M3 provides the sharpest average 90% prediction interval (187.3 m^3/h), whereas M2 provides the least sharp average prediction interval (286.3 m^3/h).

From studying the reliability and the sharpness it is not immediately clear which model should be preferred. However, this can be unravelled by calculating the skill score for each model for every prediction step and as an average for the entire prediction horizon. Table 2 shows the skill score for the generated 90% prediction intervals calculated for various prediction steps, and as an average for the maximum prediction horizon of 4 hours. Note that all 16 prediction

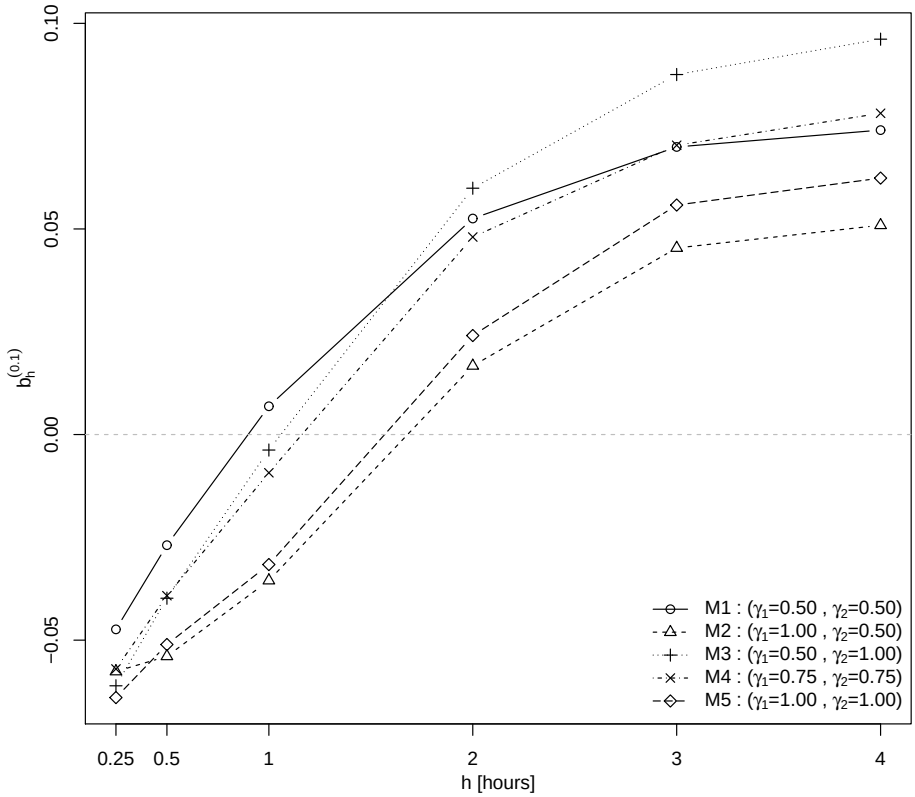

Figure 3: Reliability bias of the 90% prediction interval, as a function of the prediction horizon.

steps (every 15 minutes for 4 hours) are included in the average skill score but only 6 prediction steps are presented in Table 2. M3 is seen to perform best at prediction steps 0.25 and 0.5 hours (recalling that the smaller skill score is the preferred score), while M5 (the model with state proportional dependency for both states) performs best at larger prediction horizons up to 4 hours. Surprisingly, the most reliable model M1 is seen to perform rather poorly compared to the other models for the prediction horizons of 1h to 4h. Apparently, the sharpness for M1 is too narrow because many observations fall too far away from the lower and upper prediction bounds incurring a high penalty when calculating the skill score. When considering the average skill score for the entire prediction horizon of 4 hours, it is furthermore seen that M2 - M4 perform rather similarly, whereas M1 has a significantly higher score value.

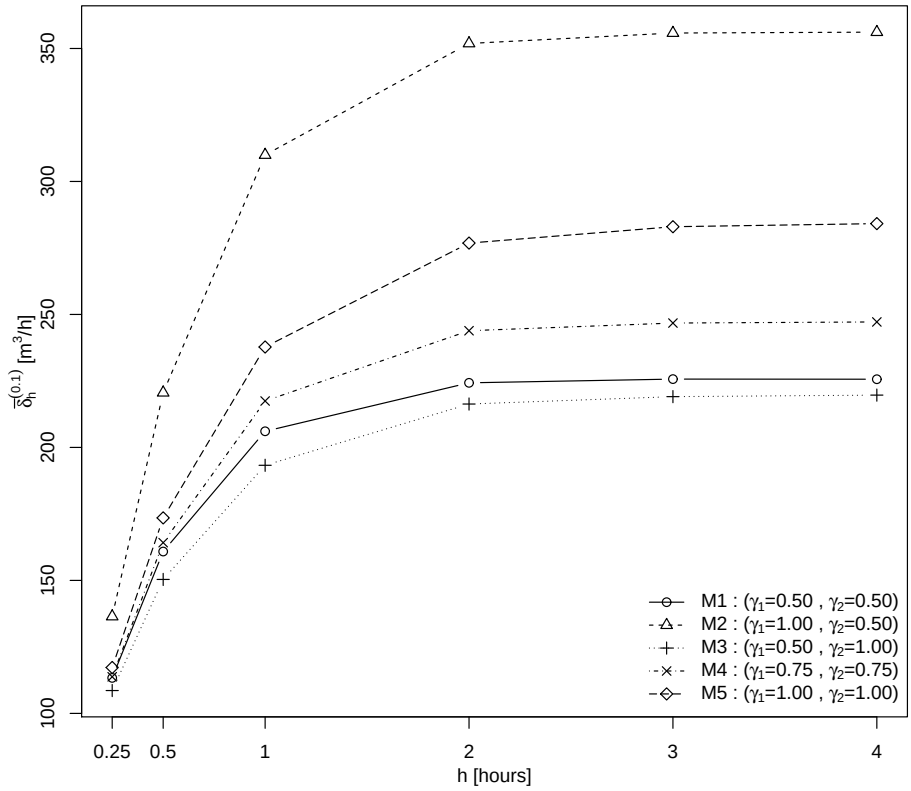

Figure 4: Sharpness for the 90% prediction intervals, as a function of the prediction horizon for all five models.

Table 2: Skill score calculated from 90% prediction intervals at several prediction steps and averaged for the entire prediction horizon of 4 h. The preferred model candidate for each prediction horizon is highlighted in bold.

			Prediction Horizon						
	γ_1	γ_2	0.25h	0.5h	1h	2h	3h	4h	Average
M1	0.50	0.50	166.0	292.7	491.2	680.8	724.7	732.7	514.7
M2	1.00	0.50	201.9	324.9	455.2	563.6	602.6	610.1	459.7
M3	0.50	1.00	**137.2**	**228.3**	391.2	603.8	675.8	691.7	454.7
M4	0.75	0.75	155.1	264.3	429.7	606.4	663.6	673.6	465.4
M5	1.00	1.00	150.4	247.1	**383.8**	**535.2**	**593.8**	**608.2**	**419.7**

4.6 Resolution analysis: conditioning on dry and wet weather periods

From a model predictive control point of view it is especially of interest to evaluate how well the models perform during wet weather periods. Separation of wet weather flow measurements from dry weather flow measurements using a rough flow threshold, i.e. wet weather interpreted as flows above 540 m^3/h and dry weather flows below, a conditional reliability is obtained as shown in Figure 5. By introducing this threshold, 90% of the flow data is catagorised in the dry weather period and the remaining 10% in the wet weather period. For shorter prediction steps, the dry weather reliability (see Figure 5a) is overestimated, whereas it is underestimated for longer prediction steps. This shift in reliability was also observed in the unconditional case seen in Figure 3, and thus emanates from dry weather periods. In wet weather periods, the underestimated reliability increases with the length of the prediction horizon; see Figure 5b. The only exception appears at the one-step prediction (0.25h), where M3 and M5 both are reliable. At the 4h prediction step the reliability bias is around 50-75%, compared with just 10% in the unconditional case. The models with $\gamma_1 = 1$ (M2, M5) are significantly less biased than the remaining models, but still underestimate the coverage by approximately 50% at the 4h prediction step. This observed discrepancy in reliability bias between the unconditional case and the wet weather periods reveals the importance of the resolution analysis, and show that the relatively low reliability bias at the 4h prediction horizon for the unconditional case is a result of the dry weather period, constituting 90% of the whole data set.

The conditional sharpness is shown in Figure 6. In dry weather periods (Fig. 6a) the sharpness is very close to the unconditional sharpness, albeit slightly more sharp. In wet weather periods (Fig. 6b), the sharpness decreases considerably, i.e. the prediction intervals are approximately twice the size in dry weather periods. It is seen that the prediction uncertainty grows rapidly during the first prediction steps and then levels out at 2h. The effect of the diffusion term is clearly identified. The models M2 and M5 are seen to be the least sharp, but both models have state proportional diffusion in the first reservoir ($\gamma_1 = 1$). In contrast, the models M1 and M3, with $\gamma_1 = 0.5$, generate the sharpest prediction intervals.

The dry weather conditional skill score for the five model candidates is seen in Table 3. It is readily seen that M3 is the preferred model candidate both at each prediction step and as an average for the entire prediction horizon. As the reliability bias was found to be close to zero at all considered prediction steps, we conclude that M3 is very useful for making 90% prediction intervals in dry weather periods.

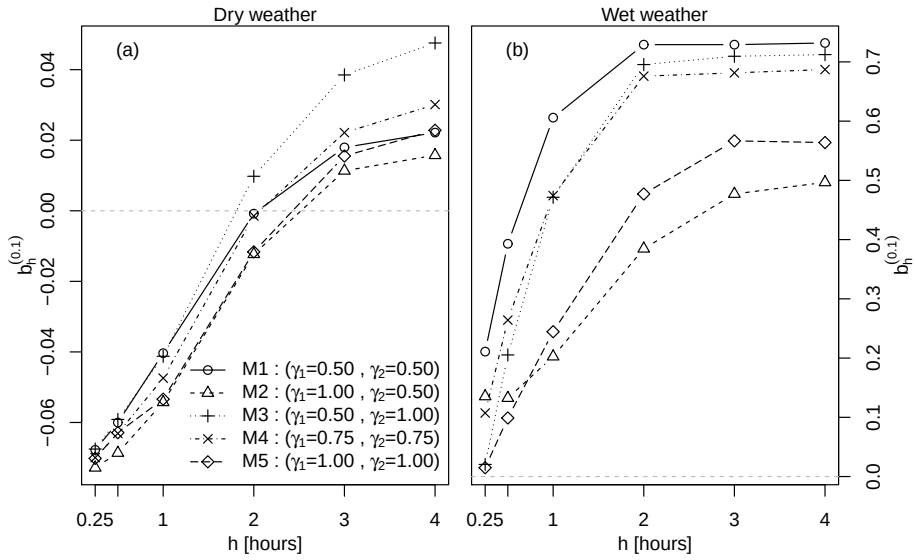

Figure 5: Reliability of the 90% prediction intervals, as a function of the prediction horizon and conditioned on the weather: (a) for dry weather periods; (b) for wet weather periods. A flow threshold of 540 m^3/h was applied for conditioning.

When conditioning on wet weather periods alone, Table 4 yields more ambiguous results. M3 is the best model for prediction steps of less then 1h, which is the same as obtained when conditioning on dry weather periods alone. However, for 1h to 4h, models M2 and M5 provide better results (lower skill score). Note the large difference in average skill score between dry and wet weather periods when comparing Table 3 and Table 4. The best model on average when

Table 3: Skill score calculated for the 90% prediction interval conditioned on dry weather periods. The preferred model candidate for each prediction horizon is highlighted in bold.

			Prediction Horizon						
	γ_1	γ_2	0.25h	0.5h	1h	2h	3h	4h	Average
M1	0.50	0.50	68.3	114.2	176.4	225.0	236.5	238.8	176.5
M2	1.00	0.50	74.6	128.5	198.9	253.3	266.8	269.2	198.6
M3	0.50	1.00	**62.2**	**101.9**	**159.5**	**214.3**	**233.2**	**237.7**	**168.1**
M4	0.75	0.75	65.7	109.6	170.8	224.4	240.9	244.4	176.0
M5	1.00	1.00	64.2	106.9	166.7	222.4	241.7	246.9	174.8

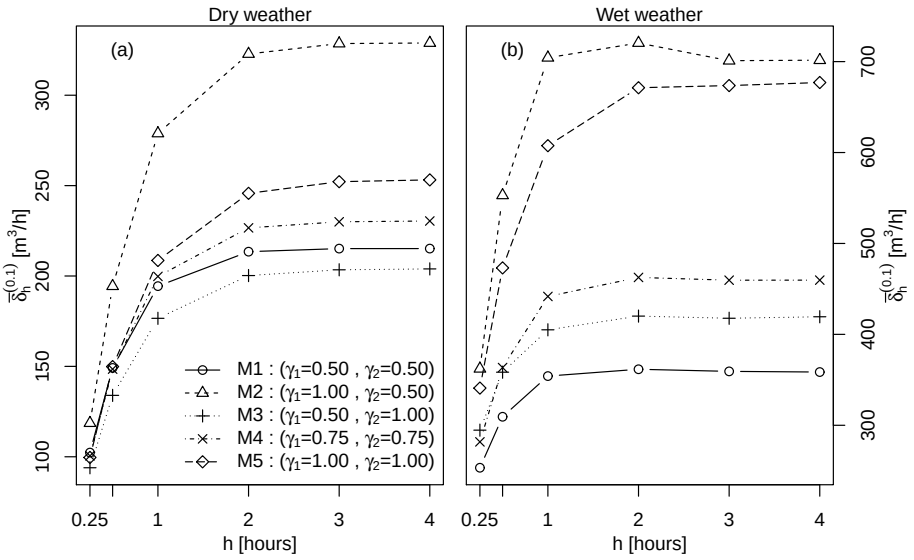

Figure 6: Sharpness of the 90% coverage, as a function of the prediction horizon and conditioned on the flow: (a) for dry weather periods ; (b) for wet weather periods. A flow threshold of 540 m^3/h was applied.

considering all prediction horizons of interest is M5, but it should be kept in mind that the reliability bias showed that none of the models are able to generate satisfactory 90% prediction intervals, and thus cannot be fully trusted when considering prediction horizons larger than one. If focusing on the one-step ahead prediction only in wet weather periods, M3 must be the preferred model; both because it was shown to be reliable and because it has the lowest skill score.

Table 4: Skill score calculated for the 90% prediction interval conditioned on wet weather periods. The preferred model candidate for each prediction horizon is highlighted in bold.

			Prediction Horizon						
	γ_1	γ_2	0.25h	0.5h	1h	2h	3h	4h	Average
M1	0.50	0.50	397.3	709.9	1243.0	1859.4	1996.8	2015.7	1370.4
M2	1.00	0.50	572.9	820.9	1044.5	**1328.0**	**1439.9**	**1456.1**	1110.4
M3	0.50	1.00	**289.1**	**489.3**	913.1	1607.8	1825.9	1874.5	1166.6
M4	0.75	0.75	372.3	631.8	1051.3	1619.2	1785.3	1815.3	1212.5
M5	1.00	1.00	365.5	583.0	**903.5**	1374.9	1547.9	1583.5	**1059.7**

The resolution study has clearly demonstrate the importance of conditioning the drainage model performance on relative weather situations; in our case rain. When considering the model performance on the whole data series, altogether it appeared as though the best model is able to provide quite reliable 90% prediction limits. However, when conditioning separately on wet weather periods it becomes clear that even the best model is unable to generate reliable prediction limits beyond 0.25h. This can primarily be ascribed to a poor rain input that does not represent the actual rainfall on the whole catchment area. If the rain input used in the models is improved by, e.g., placing rain gauges inside the catchment area or by using rain radars a different description for the diffusion term in the model would be preferred and a larger prediction horizon would probably be shown to be reliable. With more representative rain input, it is possible to extend the diffusion term by considering both the states and the rain input in its description, which would contribute to more reliable probabilistic predictions.

5 Conclusions

This study has demonstrated how simple stochastic models suitable for making interval flow predictions in urban drainage systems can be built using the grey box methodology, and the models capabilities for providing interval predictions evaluated by the performance measures: reliability, sharpness and skill score. Reliability concerns the coverage ratio of the prediction intervals that must correspond to the nominal coverage, sharpness concerns the size of the prediction interval, and finally the skill score utilises both reliability and sharpness to evaluate the prediction performance in a single score value. This is useful for model prediction comparison. Grey box models are tailored to derive the one-step prediction interval, but can, presuming a representative rain input is given and the model describes the processes well, be used to make interval predictions several time steps into the future, given that the interval predictions are reliable.

Five different grey box models, that only differed with respect to the diffusion term description, were estimated and their probabilistic prediction performance was evaluated using data from a case catchment area. A model was found that was able to predict the 90% flow prediction interval up to 4 hours ahead when all the observations were included in the study. The skill score criterion was applied to compare the prediction performance of the models and eventually to select the preferred model. However, when conditioning the model performance on wet weather periods (accounting for 10% of the whole data series), it was shown that solely the one-step prediction (15 minutes) was reliable. This can most likely be attributed to a poor rain input that does not

represent the actual rainfall on the catchment area very well. In a control context, since wet weather periods are the most important periods, more representative rain inputs and rain forecasts are needed to derive models that can reliably describe the prediction uncertainty several time steps into the future. Nevertheless, this particular case study should not detract from the power of the proposed methodology.

Acknowledgements

We appreciate the help and support of flow meter data from Spildevandcenter Avedøre I/S. The research was funded by a PhD fellowship, including DTU Informatics and DTU Environment, and by the Danish Strategic Research Council (Sustainable Energy and Environment Programme).

References

Baadsgaard M, Nielsen JN, Spliid H, Madsen H, Preisel M (1997) Estimation in stochastic differential equations with a state dependent diffusion term. *SYSID '97 - 11th IFAC symposium of system identification, IFAC*.

Barbera PL, Lanza LG, Stagi L (2002) Tipping bucket mechanical errors and their influence on rainfall statistics and extremes. *Water Science and Technology* **45**(2):1–9.

Bertrand-Krajewski, JL, Bardin JP, Mourad, M, Béranger Y (2003) Accounting for sensor calibration, data validation, measurement and sampling uncertainties in monitoring urban drainage systems. *Water Science and Technology*, **47**(2):95–102.

Breinholt A, Thordarson FÖ, Møller JK, Mikkelsen PS, Grum M, Madsen H (2011) Grey box modelling of flow in sewer systems with state dependent diffusion. Environmetrics, **22**(8):946–961.

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association* **102**(477):359–378.

Gneiting T, Balabdaoui F, Raftery AE (2007) Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B* **69**(2):243–268.

Giraldo JM, Leirens S, Díaz-Grenados MA, Rod´ríguez JP (2010) Nonlinear optimization for improving the operation of sewer systems: the Bogotá Case Study. International Environmental Modelling and Software Society (iEMSs). *2010 International Congress on Environmental Modelling and Software Modelling for Environment´s* Sake, Fifth Biennial Meeting, Ottawa, Canada.

Iacus SM (2008) *Simulation and Inference for Stochastic Differential Equations - with R Examples*. Springer series of Statistics.

Jacobsen JL, Madsen H, Harremoës P (1997) A stochastic model for two-station hydraulics exhibiting transient impact. *Water Science and Technology* **36**(5):19–26.

Jazwinski AH (2007) *Stochastic Processes and Filtering Theory*. Dover Publications, Mineola, New York, USA.

Jonsdottir H, Jacobsen, JL, and Madsen H (2001) A grey-box model describing the hydraulics in a creek. *Environmetrics* **12**:347–356.

Jonsdottir H, Madsen H, Palsson OP (2006) Parameter estimation in stochastic rainfall-runoff models. *Journal of Hydrology* **326**(1-4):379–393.

Kloeden P, Platen E (1999) *Numerical Solutions of Stochastic Differential Equations*. Springer-Verlag.

Kristensen NR, Madsen H (2003) *Continuous Time Stochastic Modeling - CTSM 2.3 - Mathematics Guide*. Technical Unversity of Denmark.

Kristensen NR, Madsen H, Jørgensen SB (2004a) Parameter estimation in stochastic grey-box models. *Automatica* **40**:225–237.

Kristensen NR, Madsen H, Jørgensen SB (2004b) A method for systematic improvement of stochastic grey-box models. *Computers and Chemical Engineering* **28**(8):1431–1449.

Limpert E, Stahel WA, Abbt M (2001) Log-normal distributions across the sciences: Keys and clues. *BioScience* **51**(5):341–352.

Madsen H (2008) *Time series analysis*. Chapman & Hall/CRC.

Mannina G, Freni G, Viviani G, Saegrov S, Hafskjold L (2006) Integrated urban water modelling with uncertainty analysis. *Water Science and Technology* **54**(6-7):379–386.

Møller JK, Nielsen HA, Madsen H (2008) Time-adaptive quantile regression. *Computational Statistics & Data Analysis* **52**:1292–1303.

Møller JK, Madsen H, Carstensen J (2011) Parameter estimation in a simple stochastic differential equation for phytoplankton modelling. *Ecological Modelling* **222**:1793–1799.

Molini A, Lanza LG, Barbera Pl (2005) The impact of tipping-bucket raingauge measurement errors on design rainfall for urban-scale applications. *Hydrological processes* **19**:1073–1088.

Nielsen HA, Madsen H (2006) Modelling the heat consumption in district heating systems using a grey-box approach. *Energy and Buildings* **38**(1):63–71.

Ocampo-Martinez C, Puig V (2010) Piece-wise linear functions-based model predictive control of large-scale sewage systems. *IET Control Theory & Applications* **4**(9):1581–1593.

Øksendal B (2003) *Stochastic differential equations - an introduction with applications*, 6th edn. Springer.

Pedersen L, Jensen NE, Christensen LE, Nielsen HA, Madsen H (2010) Quantification of the spatial variability of rainfall based on a dense network of rain gauges. *Atmospheric Research* **95**(4):441–454.

Pinson P, Nielsen HA, Møller JK, Madsen H (2007) Non-parametric probabilistic forecasts of wind power: Required properties and evaluation. *Wind Energy* **10**(6):497–516.

Puig V, Cembrano G, Romera J, Quevedo J, Aznar B, Ramoń G, Cabot J (2009) Predictive optimal control of sewer networks using CORAL tool: Application to Riera Blanca catchment in Barcelona. *Water Science and Technology* **60**(4):869–878.

Shedekar VS, King KW, Brown LC, Fausey NR, Heckel M, Harmel DR (2009) Measurement Errors in Tipping Bucket Rain Gauges under Different Rainfall Intensities and their implication to Hydrologic Models. Conf.paper, *ASABE Annual International Meeting*, June 21-24. 1–9.

Tornøe CW, Jacobsen J, Pedersen O, Hansen T, Madsen H (2006) Grey-box Modelling of Pharmacokinetic/Pharmacodynamic Systems. *Journal of Pharmacokinetics and Pharmacodynamics* **31**(5):401–417.

Vaes G, Willems P, Berlamont J (2005) Areal rainfall correction coefficients for small urban catchments. *Atmospheric Research* **77**(1-4):48–59.

Vestergaard M (1998) *Nonlinear filtering in stochastic volatility models*. Master's thesis, Technical University of Denmark. Department of Mathematical Modelling, Lyngby, Denmark.

Willems P (2001) Stochastic description of the rainfall input errors in lumped hydrological models. *Stochastic Environmental Research and risk assessment* **15**:132–152.

Willems P (2010) Parsimonious model for combined sewer overflow pollution. *Journal of Environmental Engineering* **136**(3):316–325.